摘要

我們這個研究是嘗試探究離散單元與音位之間的關係。

首先原因是因為語音本身捕捉的是連續的訊號變化，因此我們有了離散單元。然而，人類本身語音學就已經有了離散的符號 — 文字與音位。於是，我們可以試圖去比較現今語音離散表徵與人類對語音音位歸類的差異來理解這些離散表徵是否有捕捉到類似於人類發音的特性。並且，藉由語音對音位標註之間的分組，我們可以觀察這些離散表徵是否也有相似的分組特性。

最後，因為人類對音位的感知往往多於單一的語音表徵音框，因此我們可以嘗試借鑑文字處理中的次詞單位，重新編碼語音訊號再次確認這些次詞單位是否類似於人類的發音特性。

關鍵字：語音基石模型、離散單元、語音表徵、語音學

目錄

第 四 章　多個語音離散表徵與音位的關係

4.1　動機

如第三章所述，一個文字或音位往往對應到上百毫秒的語音訊號，然而單一離散單元所對應的聲音訊號為 10 或 20 毫秒，亦即同一段語音所對應的離散單元數目將比音位或文字多出許多。本章節從自然語言處理中獲取靈感，將文字處理中的分詞演算法（Tokenization）應用於離散單元序列上，使得離散單元重新組合成次詞單位（Subword Units），稱之為「聲學片段（Acoustic Piece）」，以這些由多個離散單元組成的符記（Token）[1]作為新的基本單位重新編碼語音訊號，取代原先的離散單元。為了分析聲學片段是否更接近音位的序列，在此將續用上一章節的分析方法，比對並檢驗引入次詞單位是否有機會得到更好的語音表徵，進而有機會用於無文字（Textless）架構 [51, 59, 52] 中。

4.2　相關研究

在無文字架構被提出後的約兩年後，藉助次詞單位組合離散單元的研究逐步出現。任氏（Ren）等人 [53]）最先提出聲學片段的概念，該論文比對離散單元序列及對應的文字轉寫，從中觀察到許多相似的規律重複出現，而且不限於單一語者。受此啟發，本論文首先將離散單元，透過文字處理中常用以獲得次詞單位的「句片段（SentencePiece）[8] 套件獲得新的符記 —「聲學片段」，並用於語音辨識的預訓練上。

不久，由吳氏（Wu）提出的 Wav2seq [54] 論文中，考量文字與語音的序列長

[1]指資料序列中的離散基本單位。

度差異，並基於離散單元和音位的關聯性，將離散單元視為字符（Character），嘗試將這些字符透過次詞單位組成「虛擬語言（Pseudo-language[2]）」，來幫助語音到文字的模型。在實際應用中，因為解碼器生成的目標文字序列亦是由次詞單位組成，因此該篇研究旨在讓模型在預訓練後可以快速適應下游任務。與前一篇呼應，聲學片段對語音預訓練的效果在 [14] 中被探討，此後聲學片段更被應用於縮短資料序列長度 [64] 、語音生成 [15]，或學習更穩健（Robust）的語音表徵 [13]。

近期，張氏（Chang）等人 [56] 將以分詞方法處理離散單元的流程（Pipeline）納入 ESPNet 套件 [6] 中，並在語音辨識、語音翻譯等任務中獲得了超越以往的表現，進一步證明了這個方法的效果。

4.3 文字處理中的分詞演算法

在以文字為主體的自然語言處理中，文字文本除了以單詞（Word）或字元（Character）為處理單位，更常見的作法是透過分詞演算法（Tokenization）將文本分段，以「次詞單位」構成詞彙表來重新編碼文本，用於文字模型的訓練與推理。

使用次詞單位的優點包含：

1. 固定詞彙表大小，避免未登錄詞（Out-of-vocabulary，OOV）。

2. 縮短資料序列的長度，提升訓練和推論的效率。

3. 分解單詞，捕捉更細緻的語意關係，模擬如英語中的字首（Prefix）、字尾（Suffix）等等具有特定意義的文字組合。

[2]偽語言對應之離散單元被視為「虛擬文字（Pseudo-text）」

4.3.1 常見演算法

以下介紹幾種常見的分詞方法：

位元組對編碼（Byte Pair Encoding，BPE）

位元組對編碼 [9, 10] 是一種常用的分詞方法，最初來自資料壓縮技術 [9]，後來被引入到自然語言處理領域，用以處理機器翻譯問題 [10] 。該演算法從字元開始，根據詞彙表中各個次詞單位的頻率，反覆合併常見的字元成為新的次詞單位，直到達到預定的詞彙表大小。

單詞片段（WordPiece）

WordPiece [7] 演算法由 Google 用以訓練機器翻譯系統，並在 BERT [43] 模型中被使用而廣為人知。與位元組對編碼相似，同樣是透過反覆合併的策略，但合併的依據改以機率模型取代出現頻率。

單一詞語言模型（Unigram Language Model）

單一詞語言模型 [2] 是基於語言模型的分詞方法，以機率分佈選擇次詞單位，並以最大化輸入文本的機率來為文本分段。

4.3.2 「句片段（SentencePiece）」套件

「句片段（SentencePiece）[8] 」是由 Google 開發的分詞套件，實作了前述的位元組對編碼和單一詞演算法。其優勢在於可應用於不同語言，尤其用於處理中文、日文等不使用空格分隔單詞的語言文本時，此套件大大的簡化了前處理的流程。考慮到語音訊號本身不如英語等文字，在書寫時就已經具備空格分隔單詞，因此本章節的所有次詞單位皆以句片段套件中的單一詞演算法取得。

4.4 分析方式

本章節沿用上一章節 LibriSpeech 資料集的 train-clean-100 訓練子集，以單一詞演算法取得次詞單位，並嘗試 500、1000、8000、10000、20000 五種符記種數，對每一種語音表徵和 K-平均模型的分群數，各自取得五種聲學片段文本。

比照第三章的分析方式，本章除了整體的純度與相互資訊數據外，亦同樣從聲學片段與音位的角度分別探討，藉由調整次詞單位的種類數量，探討引入次詞單位並改變符記數量，將如何影響這些符記序列與音位標註間的相關性。然而，為了避免結果呈現過於複雜，細部分析時將著重比對 500 和 1000 種次詞單位的結果變化。

由於本章節探討重點為次詞單位種數變化的影響，延續第三章的發現，後續分析將以表現最好的 HuBERT 離散表徵為主。在需要比較離散單元分群數影響時，我們將比對分群數為 50 與 100 時的差異，否則為避免數據過於複雜，離散單元的分群數預設為 50 進行細部探討。

4.5 分析結果

承繼上一個章節的分析方法，我們先將純度等數據與條件機率熱圖 $p_{y|z}(i|j)$ [3] 兩者互相對照，並以語音學排序呈現，觀察聲學片段與音位之間兩者的分佈關係。

[3] 由於共同機率分佈熱圖 p_{yz} 的數值對於觀察符記對應到音位的關係較不明顯，因此仿照 SpeechTokenizer [31]、DinoSR [19] 等論文使用 $p_{y|z}(i|j)$ 呈現。

4.5.1 由聲學片段角度探討

聲學片段數量的影響

次詞單位種數	音位純度	分群純度	音位熵	離散單元熵	PNMI
離散單元	0.5256	0.3382	3.3152	3.8681	0.4993
500	0.5574	0.0829	3.3152	6.0282	0.5357
1000	0.5744	0.0556	3.3152	6.6594	0.5466
8000	0.5957	0.0257	3.3152	8.5192	0.5729
10000	0.5955	0.0238	3.3152	8.7207	0.5750
20000	0.5921	0.0182	3.3152	9.3527	0.5820

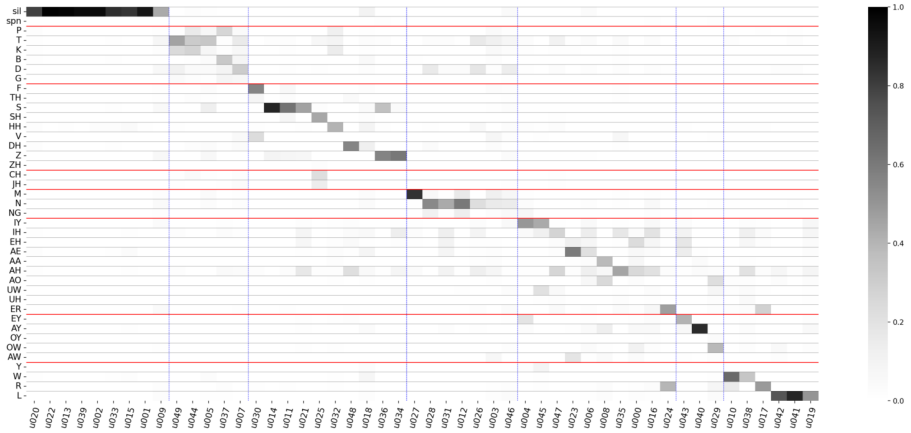
(a) 分群數 = 50

次詞單位種數	音位純度	分群純度	音位熵	離散單元熵	PNMI
離散單元	0.6097	0.2553	3.3152	4.5704	0.5786
500	0.6260	0.0972	3.3152	6.0655	0.5990
1000	0.6372	0.0631	3.3152	6.7181	0.6089
8000	0.6536	0.0237	3.3152	8.5954	0.6308
10000	0.6527	0.0219	3.3152	8.7938	0.6324
20000	0.6490	0.0173	3.3152	9.4123	0.6378

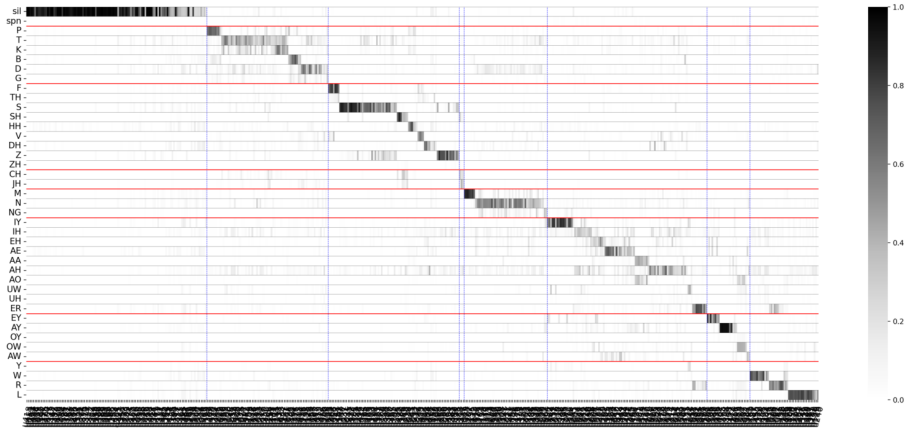
(b) 分群數 = 100

表 4.1: HuBERT 模型在不同次詞單位種類數量時的純度分析數據

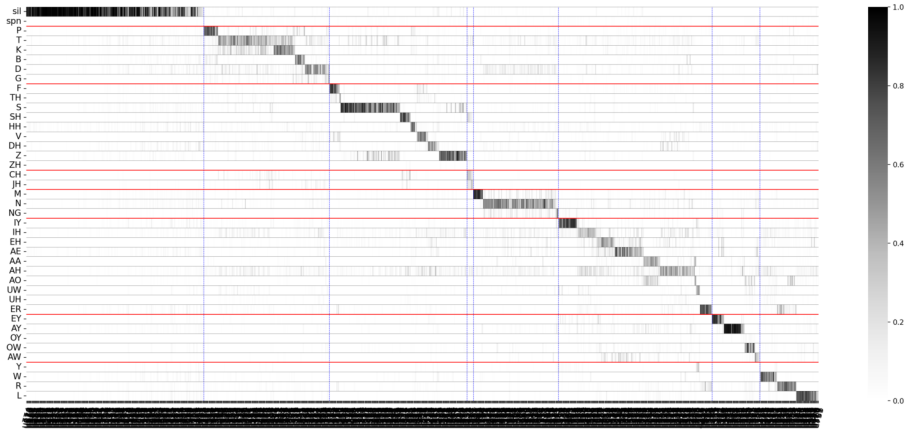
表 4.1 是 HuBERT 模型透過離散單元與不同次詞單位數量之聲學片段的純度與相互資訊數據。首先，為了觀察聲學片段數量對於機率熱圖與純度數據的影響，圖 4.1 與圖 4.2 分別以 HuBERT 表徵、分群數為 50 和 100 的離散單元為基礎，比較原始離散單元、500 和 1000 種次詞單位三種設定下，不同聲學片段數量的條件機率熱圖。從中我們可以看出，當聲學片段數量上升時，熱圖可以觀察出許多更深的色塊，也就是有更多的聲學片段可以更集中的對應到特定音位。由此可見，



(a) 離散單元



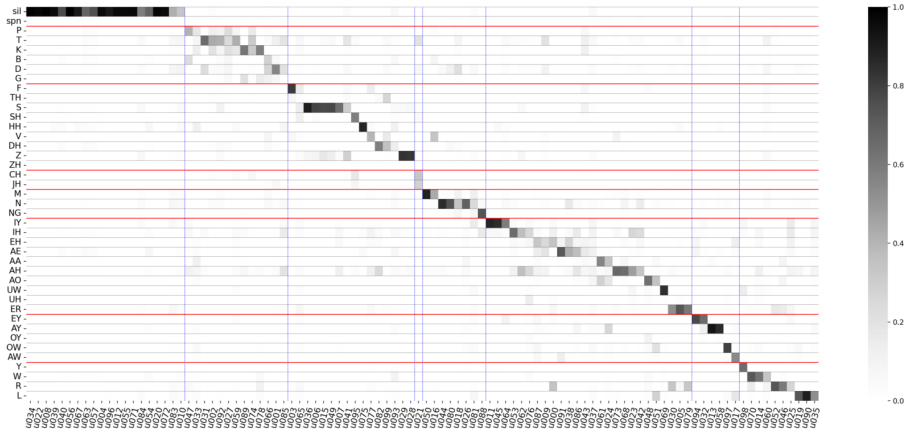
(b) 500 種次詞單位



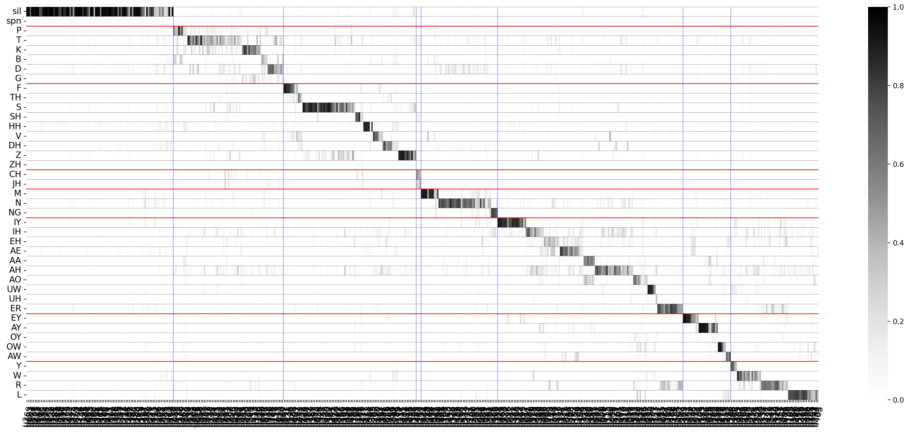
(c) 1000 種次詞單位

圖 4.1: HuBERT 表徵在 K-平均演算法使用分群數 50 後，

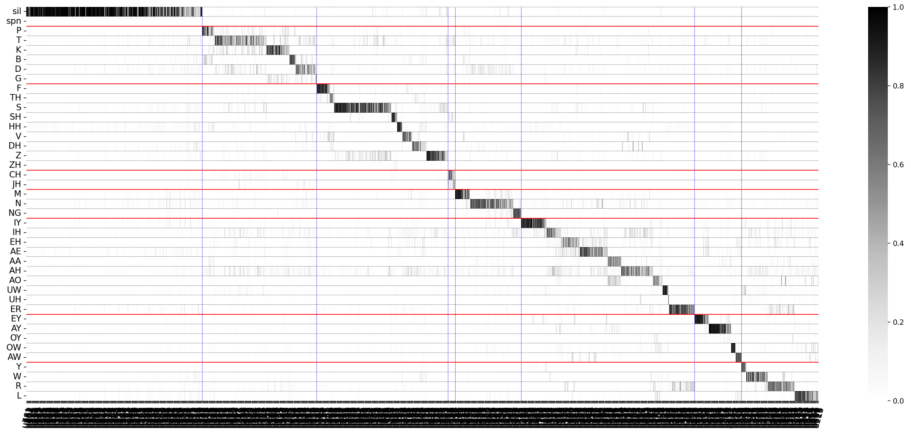
比較不同次詞單位數量的條件機率分佈 $p_{y|z}(i|j)$ 熱圖



(a) 離散單元



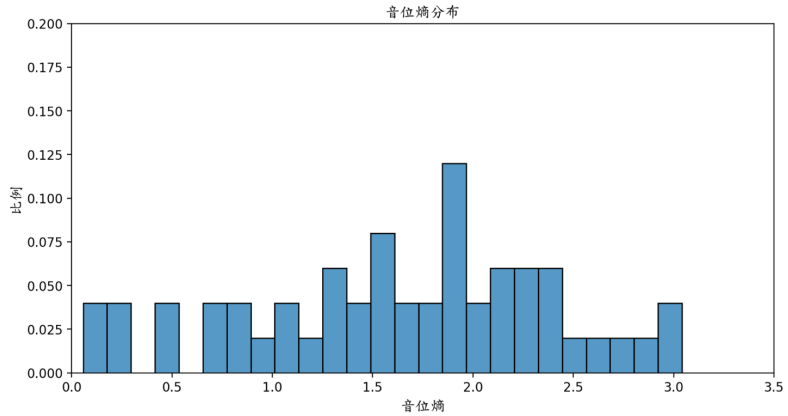
(b) 500 種次詞單位



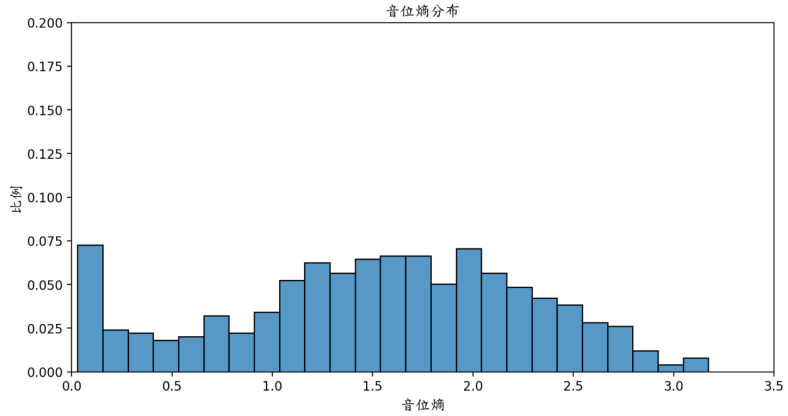
(c) 1000 種次詞單位

圖 4.2: HuBERT 表徵在 K-平均演算法使用分群數 100 後，

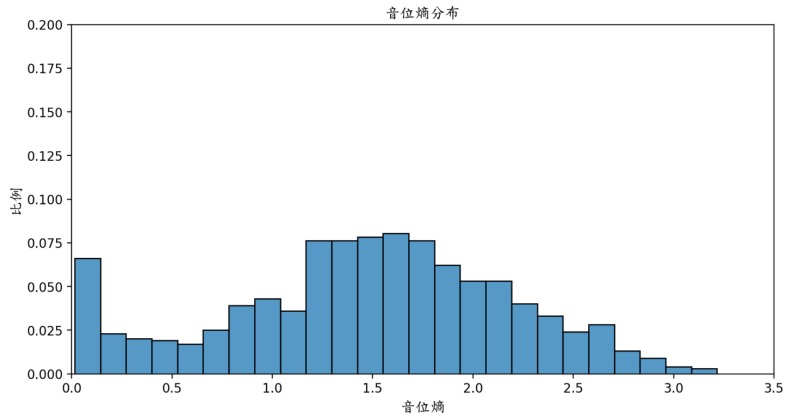
比較不同次詞單位數量的條件機率分佈 $p_{y|z}(i|j)$ 熱圖



(a) 離散單元

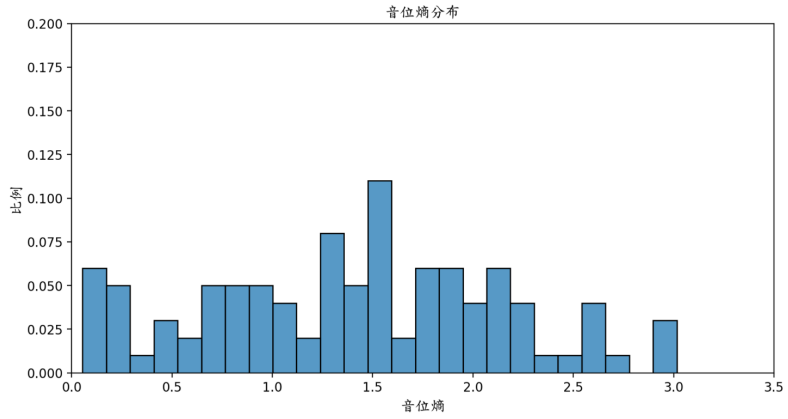


(b) 500 種次詞單位

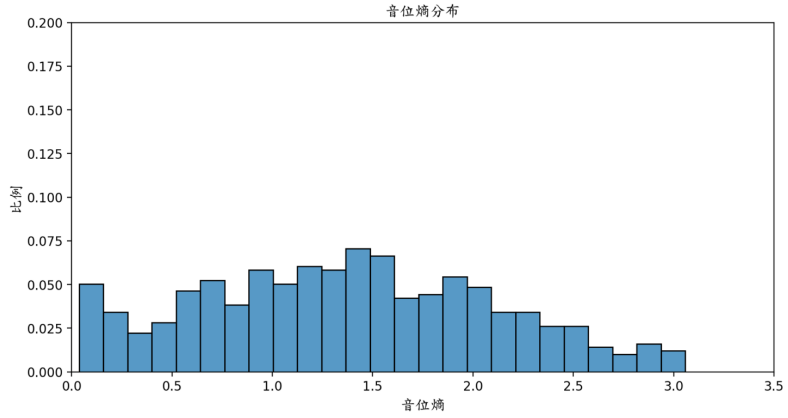


(c) 1000 種次詞單位

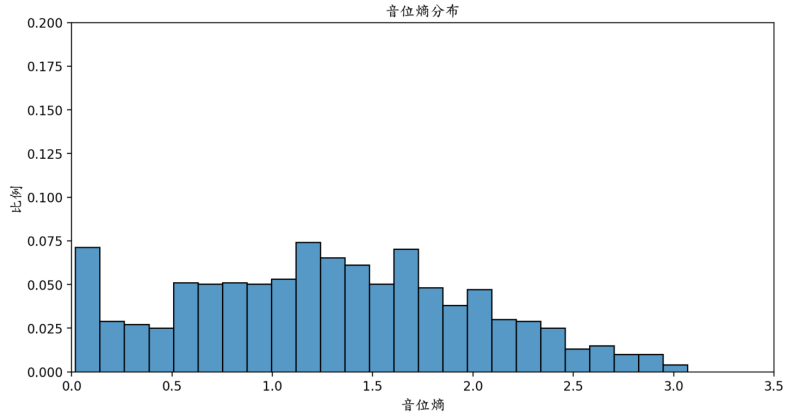
圖 4.3: HuBERT 表徵在 K-平均演算法使用分群數 50 後，比較不同次詞單位數量的音位條件熵 $H(y|z)$ 直方圖



(a) 離散單元



(b) 500 種次詞單位



(c) 1000 種次詞單位

圖 4.4: HuBERT 表徵在 K-平均演算法使用分群數 100 後，

比較不同次詞單位數量的音位條件熵 $H(y|z)$ 直方圖

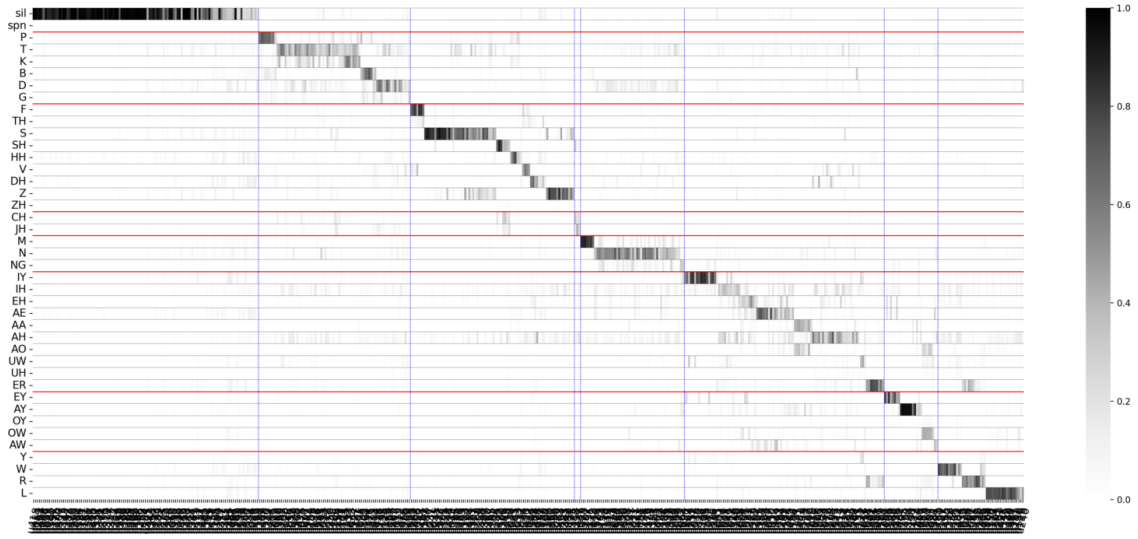
有了更多樣的符記可以區別出更細節的發音差異，使整體的純度數值有所提升；然而，機率熱圖整體也變得更加破碎，因此歸類同樣音位的效果也相對變得較不明顯。

為了確認各自聲學片段對應音位之集中狀況，我們可以考慮這些機率熱圖的條件音位熵 $H(y|z)$，以直方圖呈現來確認變化。透過觀察圖 4.3 與圖 4.4 的結果，可以確認相比第三章的離散單元，引入次詞單位確實能降低整體的條件音位熵，亦即新的符記各自能夠有更明確對應的音位，與我們從機率熱圖上所觀察到的趨勢符合。

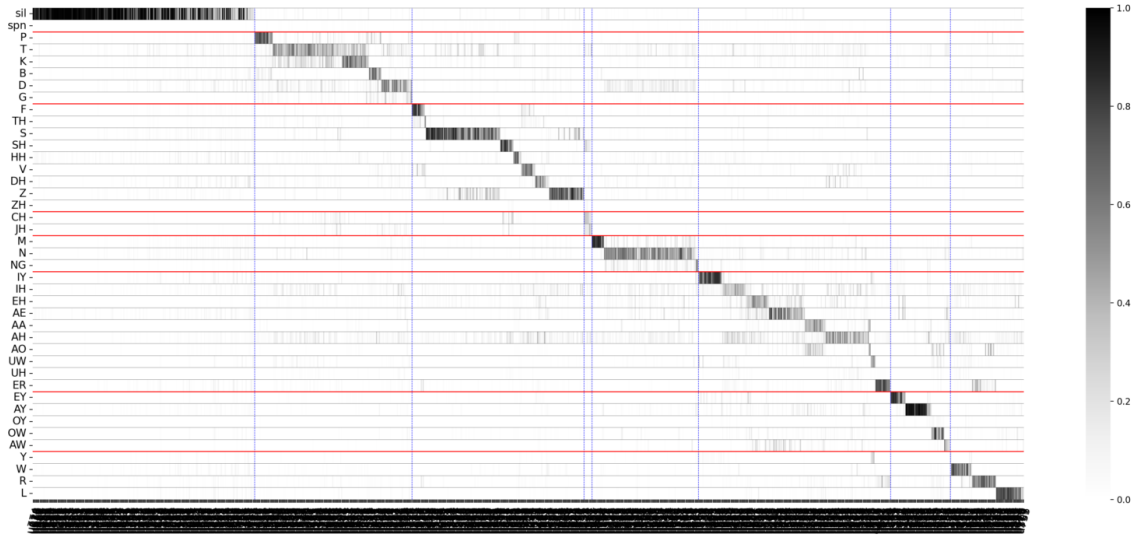
雖然改用聲學片段會使熱圖更加破碎而複雜，但除純度與相互資訊的數值變化外，觀察每個聲學片段對應之最高機率音位 $i^*(j)$ 以及它們的音位分類比例變化，也可以驗證「更多符記可以區別發音細節差異」這點。再次觀察 HuBERT 在分群數 50 時的機率熱圖（圖 4.1），圖 4.1 中三章熱圖的藍色鉛直線是每個符記在找出對應音位 $i^*(j)$ 後，按音位分類分區排序的結果。因此，比較藍色鉛直線在橫軸上各區的比例變化，可以知道有多少比例的符記能表示特定類型的發音。第三章結尾時提及過，在離散單元分群數為 50 時，由於符記數量較少，並沒有任何單元最能直接對應塞擦音音位。然而，當將這些離散單元以次詞單位進行重組後，不管在新符記種數為 500 或 1000 的機率熱圖上，都可以發現至少出現一個以上的符記得以對應到塞擦音。由此，我們驗證了引入次詞單位，對捕捉更細微的發音差異的確有所幫助。

離散單元分群數對聲學片段表現的影響

然而，儘管引入次詞單位一定程度上能幫助區別語音訊號中的細微發音差異，在語音表徵進行離散化時，K-平均演算法的分群數仍是決定這些符記捕捉語



(a) 分群數 50



(b) 分群數 100

圖 4.5: 比較同樣 500 種次詞單位的聲學片段模型，著重比較 HuBERT 表徵在 K-平均演算法使用分群數 50 與 100 的條件機率熱圖 $p_{y|z}(i|j)$ 差異

音資訊更關鍵的決定因素。圖 4.5 比較了同樣是 500 種次詞單位，K-平均演算法的離散表徵分群數選擇 50 和 100 的機率熱圖差異，不難發現分群數為 100 的機率熱圖能更加平均的對應到不同音位。然而，即便與音位的對應效果最大取決於 K-平均的分群數，但分群演算法本身相當消耗計算資源。因此當遇到運算資源限制，致使 K-平均演算法的分群數難以設置得很大時，次詞單位的引入仍舊能提升整體表現。

聲學片段對應最高機率之音位間的比較

接下來，我們比較各個聲學片段與音位之間的對應關係，亦即每個符記所對應最可能的前幾個音位之間，是否依然如離散單元那樣存在特定特徵。觀察以 HuBERT 模型、分群數 50 為基礎，分別以「離散單元」與「500 種次詞單位的聲學片段」為符記的虛擬文字文本，將對應到塞音、擦音和單元音部分的次詞單位取出觀察，將每個符記對應前五高機率的音位排名呈現在圖 4.6 中（並附上最高機率音位 $i^*(j)$ 的條件機率值 $p_{y|z}(i^*(j)|j)$），圖中上半部是離散單元，下半部則是聲學片段的結果。相互比較後可以發現，由於聲學片段的符記數量比離散單元更多，因此在維持對應音位之間相關性的同時，卻能呈現出不同音位間更細節的相關性。例如在圖 4.6a 中，上半部顯示原先以離散單元為符記時，因為只有 50 種符記，因此只能看出 T、B 與 D 比較容易和哪些其他音位比較相關，但聲學片段卻可以呈現出 P、T、B、D 等更多細節的音位關係。特別值得注意的是，圖 4.7 是對應到塞擦音的幾個聲學片段，這些對應到 CH 和 JH 兩種塞擦音的聲學片段也確實給予了同樣是塞擦音的其他音位較高的機率。

仿照第三章，藉由以音位分類作為新的標註計算純度，我們可以確認聲學片段給予同類音位較高機率的效果。然而從表 4.2 可以發現，隨著符記種數的提升，

	u049	u044	u005	u037	u007
rank 1	T	T	T	B	D
	45.88 %	31.37 %	33.04 %	32.67 %	30.73 %
rank 2	K	K	S	P	T
rank 3	D	P	K	D	B
rank 4	G	D	P	G	G
rank 5	AH	CH	Z	AH	K

	rank 1		rank 2	rank 3	rank 4	rank 5
u181	P	60.58 %	T	K	D	B
u308	P	71.21 %	T	D	K	B
u478	P	71.28 %	B	T	AH	D
u423	P	64.55 %	B	G	AH	D
u314	P	64.92 %	B	T	K	G
u346	P	67.61 %	B	G	AH	D
u401	P	56.57 %	G	D	B	T
u219	P	61.56 %	T	B	K	G
u060	P	25.55 %	B	D	AH	DH
u069	T	38.81 %	K	CH	P	D
u159	T	43.81 %	K	HH	EH	AH
u302	T	63.11 %	IH	AH	UW	K
u200	T	57.41 %	IH	UW	AH	D
u128	T	36.80 %	K	D	CH	G
u044	T	26.70 %	Z	D	S	K
u252	T	48.80 %	AH	IH	K	UW
u059	T	32.70 %	D	AH	IH	K
u028	T	23.10 %	D	N	AH	IH
u400	T	57.30 %	K	S	D	Z
u373	T	42.52 %	K	AH	IH	UW
u335	T	47.48 %	K	EH	HH	EY
u070	T	35.97 %	K	CH	D	P

	rank 1		rank 2	rank 3	rank 4	rank 5
u035	B	36.05 %	P	G	AH	IH
u204	B	46.17 %	G	D	AH	V
u394	B	63.72 %	D	G	AH	V
u196	B	61.52 %	G	D	AH	P
u371	B	75.12 %	AH	G	IH	P
u215	B	57.65 %	P	IH	G	R
u304	B	24.82 %	D	G	JH	V
u316	B	33.74 %	D	G	V	AH
u036	D	49.74 %	T	JH	G	DH
u132	D	63.13 %	JH	G	T	V
u295	D	56.94 %	G	AH	IH	T
u123	D	28.65 %	T	G	P	JH
u179	D	39.80 %	T	IH	AH	HH
u369	D	67.06 %	T	G	HH	AH
u157	D	45.86 %	T	HH	AH	IH
u009	D	14.40 %	T	N	HH	Z
u167	D	32.12 %	T	AH	IH	V
u459	D	55.26 %	G	JH	AH	B
u012	D	19.69 %	T	V	DH	N
u312	D	47.99 %	T	JH	AE	IH
u375	D	42.09 %	IH	G	T	AH
u050	D	15.14 %	N	T	sil	M

(a) 塞音

圖 4.6: HuBERT 表徵、K-平均演算法分群數 50，比較單一離散單元與使用 500 種次詞單位，依據不同音位分類比較符記各自對應的前五高音位

上半部為離散單元，下半部為聲學片段。

圖中的百分比為最高機率音位的條件機率 $p_{y|z}(i^*(j)|j)$

	u030	u014	u011	u021	u025	u032	u048	u018	u036	u034
rank 1	F	S	S	S	SH	HH	DH	DH	Z	Z
	56.54 %	87.39 %	62.24 %	46.31 %	44.75 %	40.85 %	56.04 %	12.11 %	56.40 %	60.86 %
rank 2	V	Z	SH	AH	CH	K	AH	sil	S	IH
rank 3	TH	T	Z	IH	JH	P	TH	HH	T	AH
rank 4	AH	spn	F	Z	T	T	EH	AE	AH	S
rank 5	R	HH	TH	EH	S	AE	IH	W	HH	EH

	rank 1		rank 2	rank 3	rank 4	rank 5
u093	F	67.74 %	V	TH	R	AH
u168	F	77.92 %	V	TH	R	AH
u306	F	82.32 %	TH	V	R	sil
u053	F	46.58 %	V	TH	AH	R
u344	F	84.94 %	TH	AH	R	V
u251	F	81.48 %	TH	AH	R	spn
u407	F	64.31 %	TH	AH	HH	spn
u180	S	84.29 %	AH	IH	EH	Z
u277	S	90.94 %	Z	AH	T	spn
u331	S	86.23 %	AH	IH	EH	Z
u241	S	81.50 %	AH	IH	EH	Z
u235	S	83.94 %	AH	EH	IH	Z
u047	S	54.86 %	SH	TH	F	Z
u223	S	87.27 %	Z	T	AH	spn
u476	S	93.37 %	Z	T	spn	D
u225	S	86.81 %	Z	T	D	TH
u391	S	85.69 %	Z	EH	AH	IH
u328	S	86.29 %	IH	AH	Z	EH
u147	S	60.49 %	Z	T	HH	spn
u329	S	89.34 %	Z	T	AH	spn
u434	S	61.40 %	T	P	K	Z
u015	S	41.05 %	SH	Z	TH	F
u399	S	62.53 %	T	Z	K	P

	rank 1		rank 2	rank 3	rank 4	rank 5
u023	HH	46.20 %	P	K	T	AE
u085	HH	63.62 %	P	K	T	AE
u243	HH	78.42 %	P	K	AE	EH
u032	HH	33.33 %	P	K	T	AA
u037	HH	26.92 %	P	K	R	T
u224	HH	16.69 %	sil	DH	W	AE
u030	V	44.92 %	F	TH	AH	HH
u332	V	62.37 %	AH	TH	IH	F
u422	V	53.33 %	F	AH	IH	TH
u054	V	46.19 %	TH	F	DH	AH
u043	DH	59.91 %	TH	EH	IH	AH
u106	DH	64.70 %	EH	AH	IH	EY
u208	DH	39.38 %	AH	EH	AE	TH
u311	DH	42.62 %	AH	ER	IH	IY
u097	DH	14.43 %	HH	sil	W	N
u150	DH	18.01 %	sil	HH	W	AE
u088	DH	31.73 %	TH	AH	IH	sil
u102	DH	11.75 %	sil	HH	AE	W
u100	Z	52.66 %	S	T	HH	spn
u263	Z	85.20 %	AH	IH	S	HH
u256	Z	74.83 %	IH	AH	EH	ER
u390	Z	75.49 %	IH	AH	EH	S
u299	Z	75.65 %	AH	IH	S	EH

(b) 擦音

	u004	u045	u047	u023	u006	u008	u035	u000	u016	u024
rank 1	IY	IY	IH	AE	AE	AA	AH	AH	AH	ER
	49.75 %	43.68 %	27.04 %	59.57 %	21.24 %	38.63 %	45.05 %	25.15 %	21.04 %	47.05 %
rank 2	EY	UW	AH	AW	IH	AO	IH	EH	IH	R
rank 3	NG	IH	T	EH	AH	AY	V	OW	EH	AH
rank 4	IH	Y	UW	AH	EH	AH	T	AA	ER	IH
rank 5	AY	HH	ER	AA	AO	AW	ER	AE	IY	HH

	rank 1		rank 2	rank 3	rank 4	rank 5
u087	IY	55.06 %	EY	IH	NG	N
u164	IY	65.67 %	EY	IH	NG	N
u049	IY	36.42 %	UW	Y	IH	HH
u151	IY	81.23 %	IH	HH	Y	N
u236	IY	83.38 %	IH	Y	N	HH
u038	IY	38.67 %	UW	IH	Y	UH
u212	IY	79.52 %	IH	Y	NG	HH
u326	IY	70.32 %	EY	IH	NG	N
u444	IY	87.01 %	IH	HH	N	Y
u305	IY	82.55 %	IH	HH	Y	D
u429	IY	84.49 %	IH	N	Y	HH
u182	IY	73.72 %	IH	NG	Y	N
u075	IY	33.53 %	EY	NG	AY	IH
u216	IY	76.32 %	IH	Y	NG	HH
u334	IY	79.76 %	IH	HH	Y	UW
u385	IY	58.00 %	NG	IH	Y	EY
u231	IY	15.58 %	ER	sil	D	L
u004	IH	32.68 %	AH	T	ER	EH
u074	IH	22.11 %	AH	EH	AY	N
u045	IH	35.75 %	AH	T	ER	HH
u046	IH	36.34 %	AH	ER	HH	N
u041	IH	30.11 %	IY	UW	Y	AH
u162	IH	26.01 %	AH	EH	AY	N

	rank 1		rank 2	rank 3	rank 4	rank 5
u121	AA	39.28 %	AO	AY	R	L
u171	AA	41.48 %	AO	AY	L	R
u283	AA	45.77 %	AO	AY	AH	L
u249	AA	39.58 %	AO	N	AH	T
u345	AA	40.42 %	AO	N	AH	T
u117	AA	31.29 %	AY	AO	AH	R
u260	AA	33.15 %	AH	N	AO	T
u442	AA	43.54 %	AO	AY	L	R
u120	AA	24.49 %	AH	AY	R	AO
u057	AH	58.87 %	DH	IH	IY	T
u010	AH	41.37 %	IH	V	T	ER
u005	AH	38.66 %	IH	ER	T	V
u193	AH	48.91 %	DH	IH	T	ER
u003	AH	22.09 %	EH	AY	AE	OW
u279	AH	51.51 %	EH	V	DH	M
u042	AH	22.60 %	IH	ER	N	EH
u079	AH	26.20 %	V	IH	T	B
u218	AH	45.60 %	M	EH	V	N
u194	AH	51.09 %	DH	ER	IH	IY
u297	AH	70.86 %	DH	IH	IY	M
u029	AH	20.74 %	IH	ER	IY	EH
u144	AH	42.31 %	IH	ER	V	T
u211	AH	48.92 %	V	IH	T	TH

(c) 單元音

	rank 1		rank 2	rank 3	rank 4	rank 5
u414	CH	33.07 %	SH	JH	AH	ZH
u327	JH	41.40 %	CH	AH	SH	T
u439	JH	26.49 %	CH	AH	IH	SH

圖 4.7: 對 HuBERT 分群數 50 離散單元取得 500 種次詞單位後，對應到塞擦音的聲學片段之音位條件機率排名

符記種數	音位分類純度	以音位分類標註之分群純度
離散單元	0.7006	**0.1509**
500	0.7116	0.0340
1000	**0.7186**	0.0226
8000	0.7080	0.0119
10000	0.7048	0.0113
20000	0.6929	0.0089

(a) 群數 = 50

符記種數	音位分類純度	以音位分類標註之分群純度
離散單元	**0.7584**	**0.0882**
500	0.7578	0.0326
1000	0.7576	0.0223
8000	0.7382	0.0097
10000	0.7346	0.0090
20000	0.7235	0.0074

(b) 群數 = 100

表 4.2: HuBERT 模型在不同詞表大小時的語音學類別分析數據

僅有分群數 50 時在較少符記種類時，音位分類的純度有微幅提升，多數時候符記種數的提高，伴隨的反而是音位分類純度些微的降低。由此可以推斷，次詞單位的引入雖能帶來更多樣的符記，對應音位間的關係卻在 K-平均演算法得到的離散單元已經大致抵定，次詞單位帶來的效果幾乎已經沒什麼改善的空間。其理由很可能是源自次詞單位演算法的特性，聲學片段的計算過程可以把原先代表不同種類音位的離散單元合在一起。因此，即便整體新的符記對應音位的純度有所提升，對音位分類的相關性卻低上不少。

4.5.2 由音位角度探討

考慮完聲學片段，接著我們一樣以音位的角度切入，觀察各自音位的符記分佈集中程度。圖 4.8 是對 HuBERT 模型所得的離散單元（比較分群數 50 和 100），以不同次詞單位種數取得聲學片段後，對應符記熵 $H(z|y)$ 最高與最低的排名。比對最左側直行顯示的離散單元排名，亦即第三章不引入次詞單位的結果，可以發現整體排名趨勢雖有些微變動，但對應符記最分散的音位仍以 AH、IH、T、D 為主，而最集中的亦仍然是 ZH、SH、F、EY，與上一章的觀察接近。由此可以推論，音位本身的較容易或較難以歸類的特性，在對語音表徵進行分群時就已經大致呈現；然而，即便聲學片段的演算法允許將代表不同類別音位的離散單元重新組合成新的符記，卻仍舊維持了音位本身分散程度的趨勢。因此，音位本身對應符記，不論是使用 K-平均演算法離散化獲得，或是以次詞單位重新歸類，這個分散程度的趨勢都是差不多的，音位本身的發音特徵確實是超出單一音框、影響範圍更廣的特性。

最後，我們可以將音位分類分別考慮，統計其各自的純度與相互資訊數據，與上一章節比對。對 HuBERT 分群數 50 離散單元文本以不同次詞單位種數處理

符記熵最高（HuBERT · 分群數 = 50）							符記熵最低（HuBERT · 分群數 = 50）						
次詞單位數	單元	500	1k	8k	10k	20k	次詞單位數	單元	500	1k	8k	10k	20k
# 1	spn	spn	spn	spn	spn	AH	# 1	ZH	ZH	ZH	ZH	ZH	ZH
# 2	AH	AH	T	AH	AH	spn	# 2	SH	SH	SH	SH	SH	SH
# 3	IH	T	AH	IH	IH	IH	# 3	F	F	F	EY	EY	OY
# 4	T	IH	IH	T	T	T	# 4	EY	EY	EY	W	OY	EY
# 5	D	D	D	sil	sil	sil	# 5	AA	AA	L	OY	W	W
# 6	EH	sil	sil	D	D	D	# 6	W	W	W	Y	Y	Y
# 7	TH	EH	K	EH	EH	EH	# 7	S	OW	Y	F	F	JH
# 8	HH	N	N	N	N	N	# 8	CH	L	CH	UW	UW	CH
# 9	UH	UH	EH	AE	AE	AE	# 9	IY	Y	OY	JH	JH	F
#10	G	G	G	K	DH	DH	#10	Y	CH	OW	CH	CH	UW
#11	sil	K	UH	G	K	HH	#11	OW	R	UW	L	L	OW
#12	N	TH	AE	DH	G	K	#12	Z	AO	M	OW	OW	L
#13	K	AE	TH	P	HH	G	#13	AO	M	R	IY	IY	AW
#14	AE	HH	NG	S	S	S	#14	L	UW	JH	M	M	IY
#15	P	NG	HH	TH	P	P	#15	ER	AY	AA	AO	AO	M

(a) 分群數 = 50

符記熵最高（HuBERT · 分群數 = 100）							符記熵最低（HuBERT · 分群數 = 100）						
次詞單位數	單元	500	1k	8k	10k	20k	次詞單位數	單元	500	1k	8k	10k	20k
# 1	spn	spn	spn	spn	spn	spn	# 1	SH	SH	SH	ZH	ZH	ZH
# 2	AH	AH	AH	AH	AH	AH	# 2	Y	ZH	ZH	OY	OY	OY
# 3	IH	T	T	T	T	T	# 3	ZH	Y	Y	Y	Y	Y
# 4	T	IH	sil	sil	sil	sil	# 4	F	NG	JH	SH	SH	SH
# 5	D	D	IH	IH	IH	IH	# 5	NG	F	NG	NG	NG	NG
# 6	sil	N	D	D	D	D	# 6	EY	EY	UW	UW	EY	EY
# 7	EH	sil	EH	EH	EH	EH	# 7	UW	AW	OY	F	UW	UW
# 8	HH	EH	N	S	S	S	# 8	W	UW	F	EY	F	F
# 9	UH	HH	AE	N	N	N	# 9	AW	JH	CH	JH	JH	JH
#10	AE	R	S	AE	AE	AE	#10	AY	AY	AW	CH	CH	CH
#11	K	AE	ER	K	K	Z	#11	M	OY	EY	AW	AW	AW
#12	N	TH	K	Z	Z	K	#12	AA	CH	OW	W	OW	W
#13	R	ER	HH	R	R	R	#13	OW	W	L	L	W	OW
#14	G	UH	R	HH	HH	DH	#14	CH	OW	AA	M	M	AY
#15	P	K	Z	DH	DH	HH	#15	L	AA	M	OW	L	M

(b) 分群數 = 100

圖 4.8: HuBERT 表徵、K-平均演算法分群數 50 和 100，

比較不同次詞單位種數時，符記熵最高與最低的音位排名

次詞單位數	XXX 音位純度	塞音 音位純度	擦音 音位純度	塞擦音 音位純度	鼻音 音位純度	單元音 音位純度	雙元音 音位純度	近音 音位純度
-	0.9924	0.4744	0.7033	0.6616	0.7580	0.5222	0.7813	0.8658
500	0.9931	0.5732	0.7390	0.7358	0.7745	0.5491	0.8093	0.8849
1000	0.9933	0.6002	0.7550	0.7402	0.7821	0.5658	0.8221	0.8922
8000	0.9944	0.6462	0.7949	0.8156	0.8152	0.6210	0.8607	0.9154
10000	0.9946	0.6505	0.7983	0.8209	0.8190	0.6271	0.8653	0.9181
20000	0.9951	0.6677	0.8067	0.8372	0.8313	0.6443	0.8807	0.9230

次詞單位數	XXX 分群純度	塞音 分群純度	擦音 分群純度	塞擦音 分群純度	鼻音 分群純度	單元音 分群純度	雙元音 分群純度	近音 分群純度
-	0.1733	0.2621	0.4688	0.4760	0.3633	0.3252	0.4996	0.4130
500	0.0493	0.0568	0.1063	0.0982	0.0771	0.0786	0.1308	0.1262
1000	0.0217	0.0345	0.0728	0.0876	0.0656	0.0539	0.0794	0.0977
8000	0.0149	0.0122	0.0343	0.0277	0.0321	0.0219	0.0334	0.0522
10000	0.0147	0.0117	0.0313	0.0248	0.0308	0.0198	0.0316	0.0469
20000	0.0131	0.0090	0.0252	0.0187	0.0233	0.0148	0.0219	0.0336

次詞單位數	XXX PNMI	塞音 PNMI	擦音 PNMI	塞擦音 PNMI	鼻音 PNMI	單元音 PNMI	雙元音 PNMI	近音 PNMI
-	0.8295	0.2082	0.5596	0.1065	0.3512	0.3961	0.5683	0.7110
500	0.8441	0.3531	0.6299	0.2460	0.4142	0.4389	0.6350	0.7618
1000	0.8480	0.3842	0.6477	0.2620	0.4363	0.4573	0.6682	0.7781
8000	0.8701	0.4596	0.7009	0.4226	0.5160	0.5207	0.7472	0.8185
10000	0.8746	0.4668	0.7057	0.4391	0.5253	0.5282	0.7559	0.8229
20000	0.8876	0.4947	0.7200	0.4852	0.5548	0.5505	0.7860	0.8333

圖 4.9: HuBERT 分群數 50 的離散單元，以不同符記種數取得聲學片段後，按照音位分類分開各自計算的純度與相互資訊

次詞單位數	XXX 音位純度	塞音 音位純度	擦音 音位純度	塞擦音 音位純度	鼻音 音位純度	單元音 音位純度	雙元音 音位純度	近音 音位純度
-	0.9943	0.5480	0.7535	0.6917	0.8447	0.6306	0.8273	0.8952
500	0.9948	0.6062	0.7719	0.7061	0.8490	0.6450	0.8443	0.8973
1000	0.9949	0.6415	0.7804	0.7474	0.8627	0.6554	0.8490	0.9028
8000	0.9955	0.7399	0.8193	0.8372	0.8915	0.6858	0.8797	0.9257
10000	0.9956	0.7455	0.8224	0.8446	0.8946	0.6895	0.8834	0.9275
20000	0.9960	0.7636	0.8312	0.8645	0.9030	0.7001	0.8941	0.9327

次詞單位數	XXX 分群純度	塞音 分群純度	擦音 分群純度	塞擦音 分群純度	鼻音 分群純度	單元音 分群純度	雙元音 分群純度	近音 分群純度
-	0.0855	0.1936	0.3591	0.3197	0.3325	0.2507	0.3978	0.3147
500	0.0380	0.0688	0.1288	0.1713	0.0946	0.0999	0.1791	0.1197
1000	0.0185	0.0438	0.0846	0.1380	0.0777	0.0630	0.1077	0.0830
8000	0.0093	0.0162	0.0293	0.0293	0.0294	0.0246	0.0370	0.0342
10000	0.0093	0.0143	0.0266	0.0274	0.0287	0.0221	0.0354	0.0320
20000	0.0092	0.0109	0.0206	0.0195	0.0251	0.0163	0.0270	0.0248

次詞單位數	XXX PNMI	塞音 PNMI	擦音 PNMI	塞擦音 PNMI	鼻音 PNMI	單元音 PNMI	雙元音 PNMI	近音 PNMI
-	0.8672	0.3351	0.6318	0.1998	0.5609	0.5168	0.6667	0.7734
500	0.8711	0.4236	0.6641	0.2414	0.6053	0.5406	0.7132	0.7856
1000	0.8716	0.4700	0.6773	0.3044	0.6292	0.5541	0.7260	0.8007
8000	0.8854	0.5947	0.7282	0.5106	0.7038	0.5958	0.7853	0.8391
10000	0.8880	0.6026	0.7333	0.5279	0.7113	0.6010	0.7919	0.8426
20000	0.8995	0.6308	0.7490	0.5797	0.7338	0.6170	0.8134	0.8534

圖 4.10: HuBERT 分群數 100 的離散單元，以不同符記種數取得聲學片段後，按照音位分類分開各自計算的純度與相互資訊

後，不同音位分類各自的純度與相互資訊數據以圖 4.9 呈現。由結果可以發現，隨著次詞單位種數的增加，除了原本音位純度較低的塞音在音位純度與相互資訊的提升較為明顯外，其他音位分類的音位純度與相互資訊就已經較高，因而雖然增加得不是很明顯，但整體大致仍然有所改善。比較圖 4.10，可以確認此一變化在分群數改為 100 時依然可見。

4.5.3 分析結論

藉由改以次詞單位重新組合離散單元得到聲學片段，透過符記種類數量的提升，聲學片段可以區別出語音訊號中更細節的語音差異，進而得以提升音位純度與相互資訊等數據，提高符記與音位之間的相關性。同時，次詞單位的特性雖然允許對應到不同音位的離散單元重新組合，然而如此產生的新符記，每個音位對應符記的集中或分散程度卻差異不大。因此，透過將次詞單位應用在離散單元的嘗試，結合多個離散單元重新編碼語音訊號，聲學片段在捕捉語音訊號規律上，儘管效果不如直接對語音表徵進行分群的離散單元好，卻能作為需要探索更細微語音資訊差異時，除了 K-平均演算法之外對語音訊號離散化的另一個選擇。

4.6 本章總結

本章節首先介紹了文字處理中常用的次詞單位，並嘗試對離散單元序列重新組合成聲學片段。接著，仿照第三章的分析方式，將離散單元與聲學片段互相比較，對比兩種不同符記與音位間對應關係的變化。結果顯示，無論是使用離散單元或次詞單位，儘管兩種方式在語音資訊捕捉效果上有所不同，但隨著符記種類的增加，都能獲得更加細節的語音資訊，並提升與音位之間的相關性。期望這一發現可以在未來建立語音語言模型時，除了 K-平均演算法的離散單元外，考慮與

次詞單位的演算法結合使用，以更全面和細緻的從語音訊號中提取與音位相關的語音學資訊。

第 五 章　結論與展望

5.1　研究貢獻與討論

本論文旨在細部探討和比較語音基石模型得到的離散表徵與人們理解語音的最小單位 — 音位之間的關係，藉助語音學知識為語音標註提供的分組方式，拓展純度與相互資訊給予的意義，比較共同或條件機率分佈各自的熵與純度等資訊，細部觀察機器學習到的離散表徵與音位標註之間的相似性與差異。其中，藉由分群演算法所獲得的單一離散單元，以及引入文字處理中分詞演算法重新編碼出的次詞單位 — 聲學片段，兩者都是將語音訊號離散化的方式。我們比較了離散單元與聲學片段在音位標註之間共通性的變化，探討用不同方式對語音訊號取得符記造成的影響。

首先，論文第三章介紹了與無文字架構以及語音表徵相關的分析研究，隨後簡介語音學知識中，對於不同音位之間如何按照發音特性分門別類。接著，透過純度與相互資訊定義中統計的共同機率分佈，將機率分佈從離散單元與音位兩個角度切入、比較各自的條件機率分佈特性，觀察不同模型、不同分群參數或不同音位之間是否有特定的集中或分散關係，以及不同離散表徵模型對語音訊號特性歸類的能力。結果可發現，HuBERT 作為目前無文字架構最常用的語音離散表徵模型的理由，很可能來自於它們的音位純度與相互資訊都相對較高，因而更能捕捉到語音中與內容相關的重要資訊，且同樣的趨勢在語音學分類的標註也可以被觀察到。(((寫一下細部的三方向觀察結果))) 從

其後在論文第四章，我們將離散單元以文字處理中的單一詞演算法重新分組編碼為次詞單位序列，以使得不同的離散單元之間可以重新分組成新的符記，並與第三章的結果對照比較，觀察是否在對發音特性的捕捉效果上有所變化。(((寫

一下沒什麼好處的結果)))

5.2 未來展望

希望這些對離散單元與分詞方法應用的嘗試，能幫助我們在訓練任務之前，決定哪種語音基石模型更適合作為離散編碼語音訊號的基礎。接下來，我們期望能針對常見的語音任務，特別是語音辨識和語音翻譯等內容處理相關的任務，比對離散單元促成的實際成效和分析數據之間的關係，並對這些任務中的錯誤案例進行統計和個案探討。

另外，對於如何結合語音離散單元，除了將其視為文字進行分詞演算法外，我們還可以使用其他方式對離散單元序列進行分組，以達成壓縮序列長度並使其與音位等語音內容更加一致的目標。例如，將此目標形塑為語音分段（Speech Segmentation）任務等，也是未來可以嘗試的離散單元分組方式。

最後，利用語音學分組的切入點，或許可以在未來分析離散單元或連續語音表徵時，不再僅限於參考音位或文字，還可以從語音學知識提供的相似性資訊出發，為錯誤發音修正等任務提供衡量的依據。

參 考 文 獻

[1] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," 2018.

[2] ——, "Subword regularization: Improving neural network translation models with multiple subword candidates," 2018.

[3] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[4] A. Klautau, "Arpabet and the timit alphabet," *an archived file. https://web. archive. org/web/20160603180727/http://www. laps. ufpa. br/aldebaro/papers/ak_arpabet01. pdf (Accessed Mar. 12, 2020)*, 2001.

[5] "The CMU Pronouncing Dictionary." [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict?stress=-s&in=CITE

[6] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[7] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[8] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds. Brussels, Belgium: Association for Computational Linguistics, Jan. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[9] P. Gage, "A new algorithm for data compression," *C Users J.*, vol. 12, no. 2, p. 23–38, feb 1994.

[10] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[11] Z. Ma, Z. Zheng, G. Yang, Y. Wang, C. Zhang, and X. Chen, "Pushing the Limits of Unsupervised Unit Discovery for SSL Speech Representation," in *Proc. INTERSPEECH 2023*, 2023, pp. 1269–1273.

[12] Y. Yang, F. Shen, C. Du, Z. Ma, K. Yu, D. Povey, and X. Chen, "Towards universal speech discrete tokens: A case study for asr and tts," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 401–10 405.

[13] H.-J. Chang and J. Glass, "R-spin: Efficient speaker and noise-invariant representation learning with acoustic pieces," *arXiv preprint arXiv:2311.09117*, 2023.

[14] A. Elkahky, W.-N. Hsu, P. Tomasello, T.-A. Nguyen, R. Algayres, Y. Adi, J. Copet, E. Dupoux, and A. Mohamed, "Do coarser units benefit cluster prediction-based speech pre-training?" in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.

[15] F. Shen, Y. Guo, C. Du, X. Chen, and K. Yu, "Acoustic bpe for speech generation with discrete tokens," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 746–11 750.

[16] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2021.

[17] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux, and G. Wisniewski, "Probing phoneme, language and speaker information in unsupervised speech representations," in *Proc. Interspeech 2022*, 2022, pp. 1402–1406.

[18] B. M. Abdullah, M. M. Shaik, B. Möbius, and D. Klakow, "An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech," in *Proc. INTERSPEECH 2023*, 2023, pp. 2883–2887.

[19] A. H. Liu, H.-J. Chang, M. Auli, W.-N. Hsu, and J. Glass, "Dinosr: Self-distillation and online clustering for self-supervised speech representation learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] Z. Huang, C. Meng, and T. Ko, "Repcodec: A speech representation codec for speech tokenization," *arXiv preprint arXiv:2309.00169*, 2023.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/7178964

[22] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.

[23] L. Strgar and D. Harwath, "Phoneme Segmentation Using Self-Supervised Speech Models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 1067–1073. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10022827

[24] H. Tan and M. Bansal, "Vokenization: Improving language understanding with contextualized, visual-grounded supervision," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2066–2080. [Online]. Available: https://aclanthology.org/2020.emnlp-main.162

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[30] T. Maekaku, X. Chang, Y. Fujita, L.-W. Chen, S. Watanabe, and A. Rudnicky, "Speech representation learning combining conformer cpc with deep cluster for the zerospeech challenge 2021," 2022.

[31] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," 2024.

[32] G.-T. Lin, Y.-S. Chuang, H.-L. Chung, S.-w. Yang, H.-J. Chen, S. Dong, S.-W. Li, A. Mohamed, H.-y. Lee, and L.-s. Lee, "Dual: Discrete spoken unit adaptive learning for textless spoken question answering," *arXiv preprint arXiv:2203.04911*, 2022.

[33] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," 2020.

[34] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[35] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[36] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021, publisher: IEEE.

[37] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-Task Self-Supervised Learning for Robust Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6989–6993, iSSN: 2379-190X.

[38] Y.-A. Chung and J. Glass, "Generative Pre-Training for Speech with Autoregressive Predictive Coding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3497–3501, iSSN: 2379-190X.

[39] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33.

Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[41] L. T, LiShang-Wen, and LeeHung-yi, "TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, publisher: IEEE. [Online]. Available: https://dl.acm.org/doi/10.1109/TASLP.2021.3095662

[42] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders," Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.12638v2

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers), M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013, arXiv:1301.3781 [cs]. [Online]. Available: http://arxiv.org/abs/1301.3781

[46] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019, publisher: IEEE.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://aclanthology.org/W14-4012

[49] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[50] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 27th ed. Dallas, Texas: SIL International, 2024. [Online]. Available: https://www.ethnologue.com

[51] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "Generative Spoken Language Modeling from Raw Audio," Sep. 2021, arXiv:2102.01192 [cs]. [Online]. Available: http://arxiv.org/abs/2102.01192

[52] "Textless NLP: Generating expressive speech from raw audio," Sep. 2021. [Online]. Available: https://ai.meta.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/

[53] S. Ren, S. Liu, Y. Wu, L. Zhou, and F. Wei, "Speech Pre-training with Acoustic Piece," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2648–2652. [Online]. Available: https://www.isca-archive.org/interspeech_2022/ren22_interspeech.html

[54] F. Wu, K. Kim, S. Watanabe, K. J. Han, R. McDonald, K. Q. Weinberger, and Y. Artzi, "Wav2Seq: Pre-Training Speech-to-Text Encoder-Decoder Models Using Pseudo Languages," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096988/

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp.

12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[56] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe, Y. Fujita, T. Maekaku, P. Guo, Y.-F. Cheng, P. Denisov, K. Saijo, and H.-H. Wang, "Exploring Speech Recognition, Translation, and Understanding with Discrete Speech Units: A Comparative Study," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11 481–11 485, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/10447929

[57] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong, H. Inaguma, S. Popuri, C. Wang, J. Pino, W.-N. Hsu, and A. Lee, "Speech-to-Speech Translation for a Real-world Unwritten Language," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4969–4983. [Online]. Available: https://aclanthology.org/2023.findings-acl.307

[58] H.-y. Lee, A. Mohamed, S. Watanabe, T. Sainath, K. Livescu, S.-W. Li, S.-w. Yang, and K. Kirchhoff, "Self-supervised Representation Learning for Speech Processing," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, M. Ballesteros, Y. Tsvetkov, and C. O. Alm, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 8–13. [Online]. Available: https://aclanthology.org/2022.naacl-tutorials.2

[59] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On Generative Spoken Language Modeling from Raw Audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021, place: Cambridge, MA Publisher: MIT Press. [Online]. Available: https://aclanthology.org/2021.tacl-1.79

[60] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How Much Can a Bad Teacher Benefit ASR Pre-Training?" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6533–6537. [Online]. Available: https://ieeexplore.ieee.org/document/9414460/

[61] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 2019, arXiv:1807.03748 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1807.03748

[62] X. Zhao, Q. Zhu, J. Zhang, Y. Zhou, and P. Liu, "Speech Enhancement with Multi-granularity Vector Quantization," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Oct. 2023, pp. 1937–1942, iSSN: 2640-0103. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10317485

[63] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12 644–12 652, Jun. 2023,

number: 11. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26488

[64] X. Chang, B. Yan, Y. Fujita, T. Maekaku, and S. Watanabe, "Exploration of Efficient End-to-End ASR using Discretized Input from Self-Supervised Learning," in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1399–1403. [Online]. Available: https://www.isca-archive.org/interspeech_2023/chang23b_interspeech.html

[65] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Jan. 1998, conference Name: Proceedings of the IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/726791

[66] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/

[67] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, publisher: Springer. [Online]. Available: https://doi.org/10.1007/BF02478259

[68] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, publisher: American Psychological Association. [Online]. Available: https://doi.apa.org/doi/10.1037/h0042519

[69] D. Wells, H. Tang, and K. Richmond, "Phonetic Analysis of Self-supervised Representations of English Speech," 2022, pp. 3583–3587. [Online]. Available: https://www.isca-archive.org/interspeech_2022/wells22_interspeech.html

[70] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, Jan. 1989. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0893608089900038

[71] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/323533a0

[72] D. E. Rumelhart and J. L. McClelland, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, 1987, pp. 318–362. [Online]. Available: https://ieeexplore.ieee.org/document/6302929