Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 33-42

2020 福爾摩沙臺語語音辨識比賽之初步實驗

A Preliminary Study of Formosa Speech Recognition Challenge 2020 – Taiwanese ASR

余福浩*、盧克函*、王繹崴*、張維哲*、黃偉愷*、陳冠宇*

Fu-Hao Yu, Ke-Han Lu, Yi-Wei Wang,

Wei-Zhe Chang, Wei-Kai Huang and Kuan-Yu Chen

摘要

為研究當前深度學習語音辨識模型於臺文與臺羅拼音之語音辨識任務之成效，本研究使用 2020 福爾摩沙臺語語音辨識競賽(Formosa Speech Recognition Challenge 2020, FSR-2020)所提供之臺文語音語料庫(TAT-Vol1)以及公視臺語台訓練語料，並基於 ESPnet 與 Kaldi，比較數種模型架構、訓練方法與參數於臺語語音辨識之成效。最終，在 2020 福爾摩沙臺語語音辨識競賽裡，我們的系統在臺文辨識(Track2)中取得 66.1%的錯誤率，而在臺羅拼音辨識(Track3)方面，我們所得到的錯誤率為 28.6%。

Abstract

In order to study the effectiveness of the current deep learning-based speech recognition models in the speech recognition tasks of Taiwanese Southern Min Recommended Characters and Taiwan Minnanyu Luomazi Pinyin, this study uses the corpora provided by the 2020 Formosa Speech Recognition Challenge 2020 (FSR-2020) to evalutae some neural-based ASR systems by ESPnet and Kaldi toolkits. In the end, our system achieved a 66.1% error rate in the Taiwanese Southern Min Recommended Characters recognition (Track2), and the error rate we got in the Taiwan Minnanyu Luomazi Pinyin recognition (Track3) was 28.6%.

*國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
E-mail:{M10815004, M10915010, M10915036, M10915082, M10915012, kychen}@mail.ntust.edu.tw

關鍵詞：臺文，臺羅拼音，臺語語音辨識

Keywords: Taiwanese Southern Min Recommended Characters, Taiwan Minnanyu Luomazi Pinyin, Taiwanese ASR

1. 緒論 (Introduction)

臺語(Taiwanese)又稱為臺灣話、臺灣閩南語，是臺灣地區許多人的母語（本土語言），由於過去華語政策的影響，中文(Mandarin Chinese)逐漸取代各地區母語，包括臺語、客家語以及各族之原住民語，成為臺灣目前主流的溝通語言。然而，能夠流利使用傳統母語的人數逐年下滑，以聯合國教科文組織對語言提出的世代傳承指標(Intergenerational Language Transmission)來看，臺語已處於確定瀕危(Definitely Endangered)（指小孩不再於家中如同學習母語一樣學習此語言）至嚴重瀕危（只有祖父母或更老一輩的人會說此語言；儘管父母這一代人可能理解，但不會對小孩說）之間。因此，如果小孩不再學習及使用臺語，臺語極有可能在幾十年後就成為指標中滅絕(Extinct)的語言。

近年來由於深度學習的興起，在語音辨識(Speech Recognition)領域中，深度學習模型也已逐漸取代傳統機率式模型，在許多資料集上也早已超過傳統模型，獲得相當好的辨識成果。語音辨識資料集不僅蒐集不易，語料的標註(Labeling)構是耗時費工。臺語，因為不是主流語言，要找到熟悉臺語的人士進行語料標註，就顯得更加地不容易。因此要進行臺語語音辨識器的訓練，就得面對低資源(Low-resource)甚至是無資源的問題。為維護臺語文化，北科大錄製了臺文語音語料庫(Taiwanese Across Taiwan corpus, TAT corpus) (Liao *et al.*, 2020)並舉辦了臺語語音辨識競賽(Formosa Speech Recognition Challenge 2020, FSR-2020)，比賽項目則分為三大類型，分別為將臺語語音資料辨識為繁體中文字的 Track1（即翻譯任務），以及將其辨識為臺文的 Track2 與臺灣閩南語羅馬字拼音的 Track3。在這個研究中，我們以將臺語辨識為臺文的 Track2 與臺灣閩南語羅馬字拼音的 Track3（臺羅數字調，不考慮音調部分）為研究主題，嘗試使用目前主流之端對端(End-to-end)語音辨識器進行臺語語音辨識之任務。

2. 方法 (Methods)

2.1 競賽2: 臺文辨識 (Track2: Taiwanese Southern Min Recommended Characters Recognition)

由於近年在語音辨識上基於深度模型之自動語音辨識器蓬勃發展 (Watanabe *et al.*, 2017; Gulati *et al.*, 2020; Dong *et al.*, 2018)，在許多資料集上都達到比傳統模型更好的成果，也因此這些方法逐漸取代了傳統的機率式模型（例如隱藏式馬可夫(Hidden Markov Model, HMM)模型(Bahl *et al.*, 1986)）。因此，在本研究中，我們採用基於連結時序分類(Connectionist Temporal Classification, CTC)以及注意力機制(Attention Mechanism)的混合架構(Watanabe *et al.*, 2017; Watanabe *et al.*, 2018)，探討當前語音辨識模型於臺文語音辨識的成效。

2.1.1 模型架構 (Model Architectures)

我們採用目前端對端語音辨識模型的主流架構，即基於 CTC 以及 Attention 的混合模型 (Hybrid CTC-Attention based Models) (Watanabe *et al*., 2017)，並使用端到端的語音處理工具 ESPnet(End-to-end Speech Processing Toolkit) (Watanabe *et al*., 201 8)進行實作。ESPnet 可用來完成如語音辨識、語音合成等任務，在模型方面 ESPnet 提供了很多經典的模型供開發者做使用，如：Transformer (Watanabe *et al*., 2017; Vaswani *et al*., 2017; Dong *et al*., 2018)、Conformer (Gulati *et al*., 2020)、RNN (Elman, 1990; Hochreiter & Schmidhuber, 1997; Cho *et al*., 2014)等架構；在模型後端的神經網路框架有 Pytorch 及 Chainer 兩個版本可供選擇；在語音特徵預處理方面，使用 Kaldi 來提取語音特徵；做為一個主流的開源軟體，ESPnet 也提供了許多常見語音辨識資料集之範例腳本（如：AISHELL、Librispeech、WSJ 等）。因此基於 ESPnet，可以藉由修改各個部件來完成一套新的模型架構，也可以修改範例腳本，就可以基於新的資料集訓練出一套新的語音辨識器。

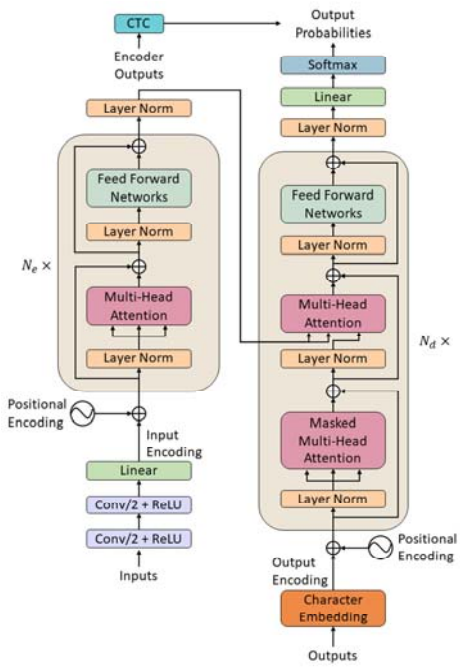

圖 1. ESPnet 中的 Transformer 模型架構圖。
[Figure 1. Model architecture of the ESPnet.]

圖 1 為我們採用的 ESPnet 架構圖，圖中左方為編碼器(Encoder)右方為解碼器(Decoder)。由架構圖可觀察出與 Speech Transformer (Dong *et al*., 2018)型相比，此模型在編碼器的輸出部分有加上 CTC 的模組。在編碼器部分，ESPnet 提供了各式各樣的神經網路模組供選擇，如：全連接神經網路或是卷積神經網路等，預設情況下 ESPnet 會使用兩層步幅為 2 的卷積層對輸入語音訊號進行降維並接著數層 Transformer 抽取出高層次的表示法，最後透過一個全連接層並計算輸出與訓練資料標籤的連結時序分類誤差，此外

依照原始 Transformer 解碼器的設計，會將編碼器的輸出傳遞至解碼器，作為解碼器中第二個多頭注意力(Multi-Head Attention)層中的鍵值(Key)及值(Value)。在解碼器輸入部分，ESPnet 提供了一般的嵌入層(Embedding)及全連接層供開發者選擇，此部分主要是用來將訓練資料標籤內的每個字元或詞轉變為向量的形式，以便進行後續的運算，最後 ESPnet 會將訓練資料標籤先進行標籤平滑化(Label Smoothing)，再接著計算解碼器輸出與訓練資料標籤的相對熵(KL Divergence)誤差，以利模型在後續進行參數更新時使用。

更明確地，ESPnet 在訓練時，會將 CTC 誤差 $\mathcal{L}^{ctc}$ 與解碼器的相對熵誤差 $\mathcal{L}^{att}$ 透過超參數 α 進行線性插值：

$$\mathcal{L} = \alpha\mathcal{L}^{ctc} + (1-\alpha)\mathcal{L}^{att} \quad (1)$$

而至於兩者之間的權重 α 為開發者可自行調整的參數，後續會利用相加過後的誤差來計算梯度並更新模型的參數；預測階段 ESPnet 使用 Beam Search 的方式進行解碼，並且在解碼階段可選擇是否要套用 RNN 語言模型來輔助模型進行預測。

2.1.2 ROVER

語音辨識模型可以透過許多方法來優化辨識的結果，例如透過提取更好的特徵或降低雜訊影響力的方式來優化模型的表現，亦或是混合多種模型的輸出結果，透過投票和重新評分的流程來決定最後的模型輸出。因此，在本研究中，我們首先訓練了許多不同架構與參數的模型後，再利用 ROVER(Recognizer Output Voting Error Reduction) (Fiscus, 1997)將模型間的輸出結果進行整合。ROVER 能夠透過投票和重新評分的過程來決定最後的輸出，進而降低預測結果的錯誤率。圖 2 為 ROVER 的示意圖。

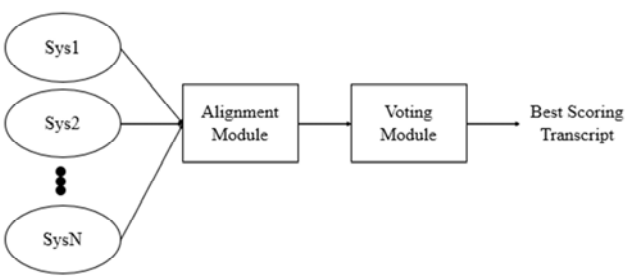

圖 2. ROVER 系統示意圖。
[Figure 2. System diagram of the Rover.]

ROVER 由兩大模組所組成，分別為對齊模組(Alignment Module)以及投票模組(Voting Module)。對齊模組會將各個辨識系統產生的結果進行對齊，接著投票模組會對每一個候選字計算一個介於 0 至 1 之間的信心分數(Confidence Score)，信心分數為 0 表示沒有任何信心，1 則表示有完全的信心。最後，投票模組有三種投票評分方式，分別為完全依靠詞頻(Word Frequency)進行投票、同時考慮詞頻與平均信心分數以及考慮詞頻與最高信心分數等方法。在本研究中，我們僅使用詞頻作為投票的依據。

2.2 競賽3: 臺羅拼音辨識 (Track3: Taiwan Minnanyu Luomazi Pinyin Recognition)

在臺羅拼音方面，本研究則嘗試使用 Kaldi 語音辨識工具中的 HMM-TDNN 混合模型，以 Tedlium 的範例為基礎修改，並使用了四連語言模型(4-gram)與 RNN 語言模型進行兩種實驗。模型架構圖如圖 3 所示。

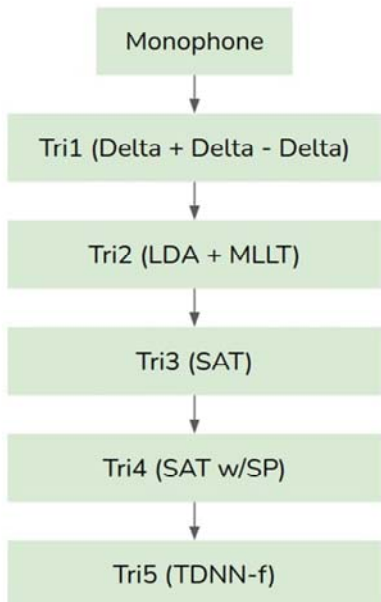

圖 3. HMM-TDNN 模型架構圖。
[Figure 3. Model architecture of the HMM-TDNN.]

3. 實驗 (Experiments)

3.1 臺文辨識 (Taiwanese Southern Min Recommended Characters Recognition)

在 Track2 的臺文辨識任務上，我們使用 ESPnet 所提供之基於 CTC 以及 Attention 的混合模型進行臺文辨識，輸入語音使用 80 維的 Fbank，模型編碼器由 12 層 Transformer 編碼器所組成（N_e=12），模型解碼器則由 6 層 Transformer 解碼器（N_d=6），多頭注意力機制數量採用 4 個 head，Transformer 中的 FFN 維度為 2048，誤差函數的超參數權重 α 為 0.3，Dropout 比率設定為 0.1，學習優化器採適應性矩估計演算法(Adam)對模型進行參數更新，並訓練 50 個 Epoch，使用 Beam Search 演算法預測時，Beam Size 設為 10，語言模型權重預設為 0.7，實驗以上述設定為基礎並嘗試各種不同的架構與設定來改善錯誤率。

在 FSR-2020 競賽期間我們將一開始收到的訓練資料(TAT-Vol1)隨機各取 10%的資料為初步實驗之開發集與測試集，進行最終實驗時我們將比賽中期釋出之 Pilot-test 測試語料加入訓練資料集中，公視臺語台訓練語料則隨機切出原始開發集與測試集之 1.5 倍左右數量之資料，加入原本的開發集與測試集中，其餘資料則做為訓練資料使用。

初步實驗使用尚未加入 Pilot-test 測試語料與公視臺語台訓練語料的資料集進行實驗，語料的統計資料如表 1 所示。

表 1. 資料集統計資訊。
[Table 1. Statistics of the dataset]

		資料數量	備註
初步實驗	訓練集	13,933	
	開發集	1,740	
	測試集	1,728	
中期實驗	訓練集	134,678	初步實驗訓練集加上公視臺語台 120,745
	開發集	1,740	
	測試集	1,728	
最終實驗	訓練集	132,142	初步實驗訓練集加入 Pilot-Test 測試語料 2,664 與公視臺語台 115,545
	開發集	4,340	初步實驗開發集加入公視臺語台 2,600
	測試集	4,328	初步實驗測試集加入公視臺語台 2,600

初步實驗集的結果如表 2 所示，使用 Transformer 架構的模型可在測試集上達到 21.7%的字元錯誤率(CER)，在使用預訓練(Pre-train)於 AISHELL-1 與 Librispeech 資料集（以字元為單位）的模型參數進行微調(Fine-tune)後，在測試集上的錯誤率可下降至 16.5%，若加上速度擾動(Speed Perturbation)可進一步下降到 15.3%，此外若使用字節對編碼(Byte Pair Encoding, BPE)為單位進行預訓練，錯誤率可再下降至 14.9%，此模型亦為繳交 Pilot-test 之模型，在 Pilot-test 所測得之錯誤率為 25.22%。

表 2. 初步實驗資料集之字元錯誤率(CER)。
[Table 2. Character error rate (CER) on the preliminary dataset.]

模型	開發集	測試集
Transformer Model	21.8	21.7
Pre-train Model (char)	16.0	16.5
+ Speed Perturbation	15.0	15.3
Pre-train Model (BPE)	**14.7**	**14.9**

於比賽中期我們嘗試將公視臺語台訓練語料先全部加入訓練集進行訓練(即中期實驗資料集)，結果如表 3 所示，由於在我們的初步實驗中，使用速度擾動與使用字節對編

碼預訓練均能有效改善錯誤率，所以這組實驗均預設地加入這兩種方式進行訓練。

表 3. 中期實驗資料集之字元錯誤率(CER)。
[Table 3. Character error rate (CER) on the intermediate dataset.]

模型	開發集	測試集
Transformer Model	16.0	16.3
Pre-train Model (BPE)	**14.8**	**15.1**

有鑑於公視臺語台訓練語料可能有較多雜音或背景音與一開始釋出的訓練語料差異較大，因此最終實驗中也將公視臺語台訓練語料加入至最終實驗的開發集與測試集之中（即最終實驗資料集）。在表 4 實驗結果中，我們嘗試使用不同架構模型（未進行預訓練且未使用語度擾動），可發現 Transformer 架構的表現比傳統 RNN 架構優秀很多，而後續推出之 Conformer 架構也能獲得比 Transformer 架構更好的結果，不過 Conformer 架構的缺點則為訓練時間會比 Transformer 架構來的更久。

表 4. 最終實驗資料集之字元錯誤率(CER)。
[Table 4. Character error rate (CER) on the final dataset.]

模型	開發集	測試集
Transformer	26.4	26.4
Conformer	**25.8**	**25.6**
RNN	29.2	29.5

接著，我們以 Transformer 模型為主，再加上以字節對編碼為單位進行預訓練於 AISHELL-1 與 Librispeech 資料集，並且在預測時調整不同語言模型權重之結果於表 5 所示。實驗結果顯示，當語言模型權重為 0.2 時可得到字元錯誤率 20.8%為最佳成果。我們也比較了 Conformer 架構與 Transformer 架構在未進行預訓練並在未使用語速擾動的辨識成效，實驗結果如表 6 所示。我們發現當權重設定為 0.1 時，可以獲得最好的結果，並且語言模型權重的影響與表 5 之結果非常接近。

表 5. Transformer 模型使用不同語言模型權重之字元錯誤率(CER)。
[Table 5. Character error rate (CER) for Transformer model with different language model weights.]

語言模型權重	開發集	測試集	初步測試集
0	20.9	21.2	9.9
0.1	**20.6**	20.9	9.6
0.2	20.7	**20.8**	**9.5**
0.3	20.8	20.9	9.6
0.5	21.6	21.7	9.9
0.7	24.5	24.3	10.6

表 6. 不同模型架構使用不同語言模型權重之字元錯誤率(CER)。
[Table 6. Character error rate (CER) for Transformer model and Conformer model with different language model weights.]

Transformer Model		
語言模型權重	開發集	測試集
0.1	**22.4**	**22.4**
0.4	24.1	23.9
0.7	26.4	26.4
Conformer Model		
語言模型權重	開發集	測試集
0.1	**23.4**	**23.2**
0.7	25.8	25.6

由於實驗訓練了多個模型，所以我們嘗試將其中三個最佳的模型以 ROVER 結合，試著得到更好的結果，使用模型包括：同時使用預訓練與語速擾動之 Transformer 模型、僅使用語速擾動之 Transformer 模型、未使用語速擾動之 Conformer 模型，經過 ROVER 產生之結果在測試集上字元錯誤率可下降至 20.3%。於 Final-test results 我們選擇繳交表五實驗中最佳的模型（即語言模型權重為 0.2）作為 X21，以及使用 ROVER 結合模型預測的結果作為 X22，序號 X21 與 X22 的系統在最後 Final-test 中分別獲得 67.7%與 66.1%的字元錯誤率。

3.2 臺羅拼音辨識 (Taiwan Minnanyu Luomazi Pinyin Recognition)

在臺羅拼音辨識的任務上，我們將比賽提供的資料切割成訓練集、開發集與測試集，其統計資料如表 7 所示；語音辨識模型採用 Kaldi 之 HMM-TDNN 混合模型，並使用四連語言模型(4-gram)與 RNN 語言模型，兩種實驗結果如表 8 所示。實驗結果中， Baseline 模型為使用主辦單位提供之基礎系統進行訓練。在 Final-test 中，我們系統的字元錯誤率為 28.6%，與訓練時的成果有一段差距。可能的原因是因為 Track3 中無法使用公視臺語台、民視電視劇資料集，只使用了 TAT-Vol1 進行訓練，而比賽最終測試集包含合成噪音的語音，實驗中只有乾淨的音檔，並且沒有對噪音進行額外的處理。因此在資料量少，且沒有複雜的前處理下，我們在 Final-test 中並沒有獲得特別突出的成果。

表 7. 資料集統計資訊。
[Table 7. Statistics of the dataset.]

	資料數量
訓練集	18,101
開發集	2,282
測試集	2,218

表 8. 臺羅拼音辨識之字元錯誤率(CER)。
[Table 8. Character error rate (CER) for Track3.]

模型	開發集	測試集
Baseline	4.97	6.13
N-gram LM	2.77	3.75
RNN LM	2.85	3.91

4. 結論 (Conclusions)

在臺文辨識的 Final-test 上，我們所得到的錯誤率為 66.1%(X22)、67.7%(X21)，與測試在自行分割的測試集上相差許多；臺羅拼音辨識方面，我們所得到的錯誤率為 28.6%。後續觀察 Final-test 的音檔可發現，也許是因為 Final-test 中的音檔與訓練資料集形式差異較大所導致，先前所釋出音檔多為特地錄製的單人語料與公視電視台語料，而 Final-test 音檔不但有廣播、多人對話，有些語料還有噪音或是聲音忽大忽小的問題，且每段的說話時間也較長，由於未考慮這些特殊狀況來進行預處理，模型在預測上容易無法成功辨識的情形，所以導致結果不是很好。

參考文獻 (References)

Bahl, L., Brown, P., De Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86.)*, *11*, 49-52. https://doi.org/10.1109/ICASSP.1986.1169179

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., …Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In arXiv preprint arXiv:1406.1078

Dong, L., Xu, S., & Xu, B. (2018).Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5884-5888. https://doi.org/10.1109/ICASSP.2018.8462506

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211. https://doi.org/10.1016/0364-0213(90)90002-E

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 347-354. https://doi.org/10.1109/ASRU.1997.659110

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., …Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In arXiv preprint arXiv:2005.08100

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Liao, Y.-F., Chang, C.-Y., Tiun, H.-K.., Su, H.-L., Khoo, H.-L., Tsay, J. S., Tan, L.-K., Kang, P., Thiann, T.-g., Iunn, U.-G., Yang, J.-H.,…Liang, C.-N. (2020). Formosa Speech Recognition Challenge 2020 and Taiwanese Across Taiwan Corpus. In *Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA '20)*, 65-70. https://doi.org/10.1109/O-COCOSDA50338.2020.9295019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., …Polosukhin, I. (2017). Attention is all you need. In arXiv preprint arXiv:1706.03762

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253. https://doi.org/10.1109/JSTSP.2017.2763455

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., …Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. In arXiv preprint arXiv:1804.00015