Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 47-56

中英文語碼轉換語音合成系統開發

Development of Mandarin-English Code-switching Speech Synthesis System

練欣柔*、黃立宇*、陳嘉平*

Hsin-Jou Lien, Li-Yu Huang, and Chia-Ping Chen

摘要

本論文提出中英文語碼轉換語音合成系統。為了使系統可專注於學習不同語言間的內容，利用已統一語者風格的多語言人工資料集進行訓練。之後在合成器中加入語言向量，以增加系統對多語言的掌握。此外對輸入的中、英文分別進行不同的前處理，將中文進行斷詞且轉為漢語拼音，藉此增加語音的自然度，且減輕學習時的複雜度，也透過數字正規化判斷句子中的阿拉伯數字，是否需要加上數字單位。英文部份則對複雜的頭字語進行讀音判斷與轉換。

Abstract

In this paper, the Mandarin-English codeswitching speech synthesis system has been proposed. To focus on learning the content information between two languages, the training dataset is multilingual artificial dataset whose speaker style is unified. Adding language embedding into the system helps it be more adaptive to multilingual dataset. Besides, text preprocessing is applied and be used in different way which depends on the languages. Word segmentation and text-to-pinyin are the text preprocessing for Mandarin, which not only improves the fluency but also reduces the learning complexity. Number normalization decides whether the arabic numerals in sentence needs to add the digits. The preprocessing for English is acronym conversion which decides the pronunciation of acronym.

關鍵詞：語音合成、語碼轉換、資料前處理
Keywords: Speech Synthesize, Codeswitching, Text Preprocessing

* 國立中山大學資訊工程學系
Department of Computer Science and Engineering, National Sun Yat-sen University
E-mail: {m103040105, m093040070}@nsysu.edu.tw; cpchen@cse.nsysu.edu.tw

1. 緒論 (Introduction)

語碼轉換（Code-switching）是指在一句話或多句話裡，含有一種以上的語言被交替使用，這種情況在現今社會中十分常見，為因應這種趨勢，語音合成系統也朝著多語言（Multilingual）的方向發展。現今的語料多為單語言，較少有同一語者的多語言語料，這導致語碼轉換在訓練時會遇到許多問題，像是語者無法合成非母語的語句，亦或是語者隨著句子語言轉換而改變的狀況。為解決上述問題，我們參考任意語者風格中英文語音合成系統(Wang, 2021)，做為我們的資料生成模型，給予系統一個參考音檔，其可生成與參考音檔相同語者風格的聲音訊號，藉此系統統一多語言資料集的語者風格，以建置多語言語音合成系統。

在本文中，使用 FastSpeech2 (Ren *et al*., 2020) 做為合成器， 將編碼器與解碼器改為(Gulati *et al*., 2020) 所提出的 Conformer 架構，聲碼器使用 HiFi-GAN (Kong *et al.*, 2020)。此外為增加系統對於多語言的掌握，於合成器中加上語言向量（language embedding），並為句子依中、英文編上語言 ID（language ID），而語碼轉換的句子無法直接以單一語言 ID 表示，對此在實驗中進行了處理。

我們發現因中文數量龐大、字詞讀音多變，因此將中文字轉換為漢語拼音，降低系統學習時的複雜度。此外也發現交雜在中文句子中的阿拉伯數字，有需要數字單位與否的問題，於是對此進行正規化。而在英文中經常使用的頭字語，是指將一句話或較長的名詞，縮寫成連續大寫字母，其發音分為字母讀音或視為新單字，要系統完整學習所有的英文頭字語是較為困難的，我們創建頭字語字典，以便進行分類讀音方式，並進行轉換，我們透過對文本進行資料前處理，以降低複雜度，提升中文、英文語音合成之正確率。

論文之其餘章節安排如下，章節二：研究方法描述系統架構、改進方法及文字前處理；章節三：實驗設置描述資料集與模型參數設定；章節四：實驗結果對基礎架構與改進後的系統進行比較；章節五：總結我們系統的優點和未來的改進方向。

2. 研究方法 (Research Methods)

以 Conformer-FastSpeech2 加上語言向量作為模型架構，為了提升中、英文語音合成的品質，分別對輸入的中文和英文文本做不同的前處理。在中文方面，使用中文斷詞、文字轉拼音與數字正規化，英文則執行縮寫讀法判斷與轉換，並且針對語碼轉換做語言 ID 編碼，與因中、英文語速差異進行的調整。

2.1 多語言語音合成系統 (Multilingual Speech Synthesis System)

在本文中， 使用 FastSpeech2 (Ren *et al*., 2020) 做為合成器， 聲碼器使用 HiFi-GAN (Kong *et al*., 2020)。

FastSpeech2 是一個非自迴歸(Non-autoregressive)的模型，可用更短的時間合成出與自迴歸 (Auto-regressive) 模型相同品質的語音。架構中的編碼器和解碼器使用

Transformer 架構，在我們系統中將 Transformer 改為 Conformer (Gulati *et al*., 2020)，並命名為 Conformer-FastSpeech2（CFS2）。Conformer 結合了 Transformer 和卷積模組 (Convolution module) 以增強效果，其網路包含前饋神經網路（Feed Forward Module）、多頭自注意力機制（Multi-Head Attention Module)、卷積模組、層正歸化（Layer Normalization)。系統架構如圖 1，而圖右半邊則為 Conformer 的架構。

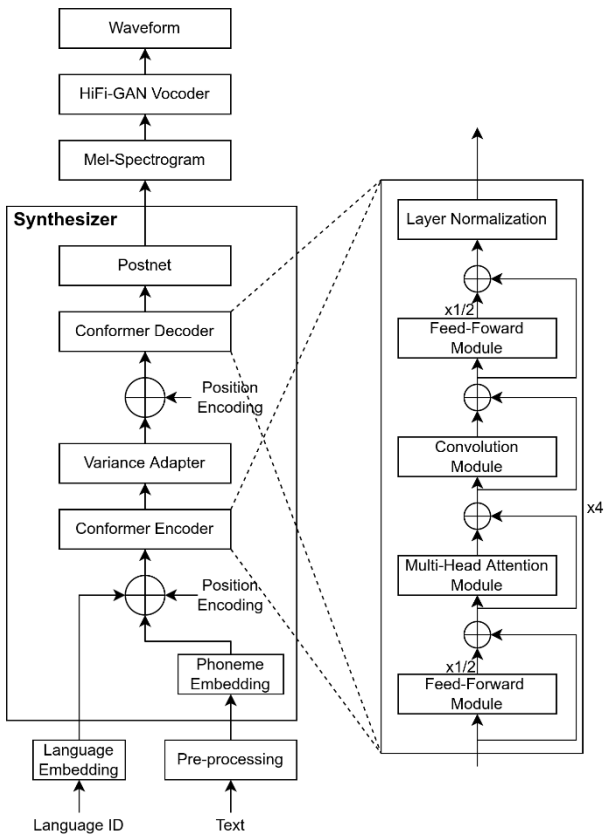

圖 1. Conformer-FastSpeech2 系統架構圖，基於 FastSpeech2 加入語言向量。右半邊為 Conformer 的架構，其基於 Transformer 架構再加上卷積模組以增強效果。

[Figure 1. The architecture of the Conformer-FastSpeech2 is based on the backbone of FastSpeech2 and combined with language embedding. The right handside of the figure is the architecture of Conformer which is associated with Transformer and convolution module to enhance feature extraction.]

在架構中加入語言向量， 並將其和 phoneme embedding 串接在一起做為編碼器的輸入。藉此提升系統合成多語言的表現，另外，依照資料的語言給予編號，稱為語言 ID，使用 0 和 1 為 language ID 分別表示英文與中文。

2.2 中文資料前處理 (Mandarin Data Preprocessing)

由於人們在交談時是有些微停頓的，為了讓系統學習語音這些細節， 我們首先使用了一個 python 工具名為 Jieba (Sun, 2012) 進行中文斷詞（Word segmentation），當中共有四種不同的斷詞模式，實驗中使用預設的精確模式，利用將符號置於斷詞處，以表示語句中的停頓，進而提升語音的自然度。此外 Jieba 工具可自行匯入符合使用者需求的字典，實驗中將 CKIP team (Ma & Chen, 2004) 的字典匯入，以提升斷詞的準確度，也將 CLMAD (Bai *et al*., 2018) 整理成另一份擴充字典，當系統應用於特殊領域時可匯入。

然而因為中文字本身數量龐大、字詞讀音多變，要系統學習所有的字詞是過於複雜的，因此不可直接將其作為輸入。於是我們利用 pypinyin 將中文字轉換成漢語拼音，其為一個 grapheme-to-phoneme（G2P）的 python 工具，以英文表示拼音，並用數字表示聲調（Tone），藉由此拼音組合的轉換簡單化中文的表示，使系統可以用較簡單的方式學習中文的發音。表 1 為中文斷詞及文字轉拼音的範例。

表 1. 中文資料前處理。先對文本進行中文斷詞，再將其轉換為漢語拼音。
[Table 1. Mandarin Data Preprocessing. Word segmentation would be applied before text-to-pypinyin.]

前處理方法	文本狀態
原始文本	明天不會下雨
中文斷詞	明天* 不會* 下雨
文字轉拼音	ming2 tian1 * bu4 hui4 * xia4 yu3*。.

此外我們還發現，當阿拉伯數字若在中文句子中時，會有是否需要唸出數字單位的差異，數字單位是指個、十、百、千、萬等。因此我們參考 [1]Chinese Text Normalization 作為基礎概念，其做法為將數字的常用情況進行分類，並以 Regular Expression 對數字找出相對應的模式，再判斷是否需加上數字單位，然而我們對模式內容進行修改，使其更貼近我們所需，共有五大種模式，表 2 為各模式的例子及正規化後的結果。

2.3 英文資料前處理 (English Data Preprocessing)

英文的頭字語可細分為 acronym 和 initialism，兩者的差異是縮寫後的單字該如何發音。acronym 指將縮寫後的單字讀為一個新的詞，例如：NASA 會讀做“na-suh”，FOMO 讀做“fow-mow”，而 initialism 則是指在發音上只念字母的讀音，而非視為一個新的詞，像是 FBI、NBA、BBC 等。然而由於頭字語為 acronym 或是 initialism，較難單純以文字進行分類，這導致系統難以學習，因此我們收集大量的頭字語，自行建立了一個頭字語字典，當輸入的文本含有全大寫的英文時，搜尋字典確認此輸入是否為 initialism，若是，則將字母轉換為相似讀音，以增加合成的正確性，若非則不做更改，舉例來說，當 BBC 經確認是 initialism ，會轉換為“bee bee ci”，FBI 則會轉換為“ef bee I”。

[1] https://github.com/speechio/chinese_text_normalization

表 2. 輸入的文本以 Regular Expression 找出相對應的模式，判斷是否要加上數字單位或其他處理。

[Table 2. Use Regular Expression to check the pattern of the text, and decides whether it requires additional number units.]

模式名稱	範例文本	正規化結果
Date	1986 年 8 月 18 日	一九八六年八月十八日
	1997/9/15	一九九七年九月十五日
Money	19588 元	一萬九千五百八十八元
Phone 手機	0919114115	零九一九一一四一一五
Phone 市話	02-2720-8889	零二二七二零八八八九
percentage	62%	百分之六十二
cardinal 量詞	1999 個蘋果	一千九百九十九個蘋果
	130 顆球	一百三十顆球
	124000 瓶水	十二萬四千瓶水
cardinal 編號	cardinal 編號學號是 103040100	學號是一零三零四零一零零
cardinal 純數	175.5 公分	一百七十五點五公分

2.4 針對語碼轉換之處理 (Process for Code-switching)

在訓練階段，使用英文和中文兩種語言 ID 進行語言向量。然而在合成階段，若輸入為語碼轉換的文本，無法單純以中文或英文予以編號。為此設立編定語言 ID 於語碼轉換文本之方法，如圖 3 所示，首先依語言分段輸入的文本，計算各分段的字元長度，藉由相對位置予以對應的語言 ID 且進行語言向量。分段後的文本分別進行資料前處理，再進行音素向量（phoneme embedding）作為編碼器的輸入。最後將編碼器輸出的隱藏特徵序列（hidden state sequence），和語言向量的輸出相加，獲得新的隱藏特徵序列進行後續的訓練。

由於中、英文資料集的語速差異，導致系統在合成語碼轉換之句子時，會有英文部份語速較快而感到不自然的問題。FastSpeech2 架構中的 Length Regulator，有一參數 α 可調整 duration predictor 輸出的時長（duration）大小，藉此改變梅爾頻譜圖的隱藏特徵序列長度，α 預設為 1。若 α= 1.5，表示將時長序列乘上 1.5 倍，進而使隱藏特徵序列拉長 1.5 倍，即為放慢速度。搭配語言 ID，即可透過相對位置單獨調整英文的速度。兩者差異如圖 2。左圖為無搭配語言 ID，針對整個序列進行調整。右圖則為單獨對英文進行調整，將英文部份的時長與 α 相乘，四捨五入，獲得新的時長序列。

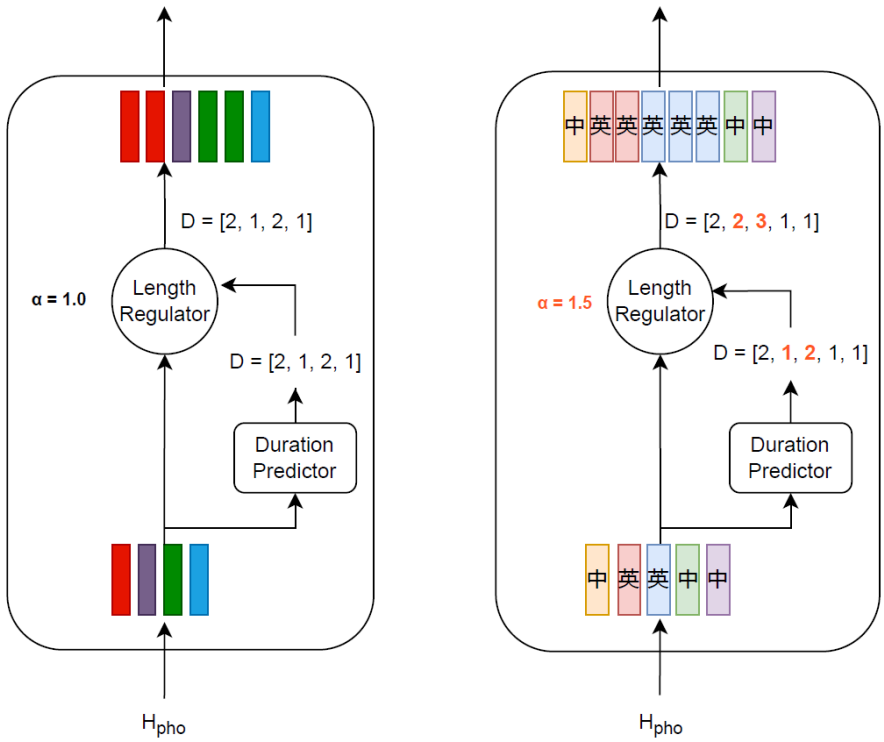

圖 2. Length Regulator 的作法。以 Length Regulator 中之參數 α 調整隱藏特徵序列長度架構圖。D 表示時長（duration），H_{pho} 表示 phoneme 的隱藏特徵，H_{mel} 為梅爾頻譜圖的隱藏特徵。右側為依語言 ID 選取要調整的時長元素，再將元素乘上 α 後四捨五入，得到新的時長以調整序列長度，左側則為對全部序列進行調整。

[Figure 2. Length Regulator. Use the parameter α in Length Regulator to adjust the length of the hidden state sequence. D denotes duration. H_{pho} denotes the phoneme hidden state. H_{mel} denotes the mel-spectrogram hidden state. The right handside of the figure shows that the specific duration is decided by the language ID. The duration elements multiply α and round it to get a new duration sequence. The left hand side of the figure adjusts all sequence.]

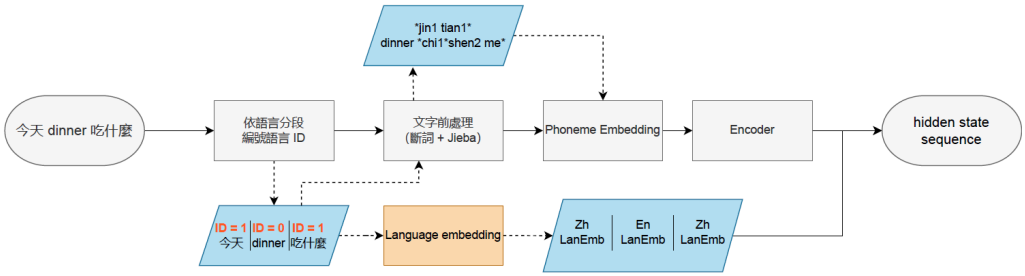

圖 3. 語碼轉換語言向量流程圖，LanEmb 表示語言向量。將輸入文本依語言分段並編號語言 ID，每段依序進行資料前處理、音素向量，將結果做為編碼器的輸入。對語言 ID 進行語言向量，將輸出與編碼器的輸出相加。

[Figure 3. Flow chart of the code-switching language embedding. LanEmb denotes language embedding. The input would be categorized by its language and the corresponding language ID would be given. After that, the result of the data preprocessing and phoneme embedding would be the input of the encoder. Finally, the output of the encoder would merge with the result of language embedding.]

3. 實驗設置 (Experiments)

在實驗中的資料集分為原始資料集，以及利用資料生成系統所生成的人工資料集，此外使用 ESPnet2 (Watanabe *et al*., 2018) 做為開發工具協助開發。

3.1 資料集 (Datasets)

- 原始資料集：使用的資料集包含中文語料 AISHELL3 (Shi *et al*., 2020) 及英文語料 VCTK (Yamagishi *et al*., 2019) ，在實驗中發現無需使用全部的資料，即可訓練出一個品質相當的系統，減少資料量亦可減少整體訓練時間，因此各選取了 30 名語者的資料做為我們實驗用的資料集，時長約為整體資料集的四分之一，並命名為 AISHELL3-thirty 和 VCTK-thirty，資料集的詳細資訊如表 3 所示。
- 人工資料集: 參考任一語者風格中英文語音合成系統(Wang, 2021)，作為我們的資料生成系統。選用 AISHELL3 資料集中的一個音檔作為參考音檔，並使用 AISHELL3-thirty 和 VCTK-thirty 的文本作為生成資料時的文本，藉此生成與參考音檔相同語者風格的多語言資料集，將其稱為 Generated-multi，共 25,362 筆音檔，共 15.6 小時，如表 3 所示。

3.2 訓練設定 (Implementation details)

本文使用 ESPnet2 (Watanabe *et al*., 2018)作為開發的工具。CFS2 的訓練集為多語言的 Generated-multi，架構中的 Conformer 編碼器和解碼器 kernel size 分別為 7 及 31，padding

為 3 與 15，優化器（Optimizer）使用 Adam (Kingma & Ba, 2014)，學習率（Learning rate）設定為 1。因為我們的系統為多語言的語音合成系統，為了使聲碼器可將多語言的梅爾頻譜圖轉為聲音訊號，HiFi-GAN 聲碼器利用 AISHELL3-thirty 和 VCTKthirty 資料集進行訓練， 批量大小（Batch size）設定為 32，使用 Adam 作為優化器，學習率設定為 0.0002。

表 3. 資料集詳細資訊。包含選取三十位語者的 VCTK-thirty 和 AISHLLE3-thirty，及生成資料集 Generate-multi。

[Table 3. The details of the dataset contain VCTK-thirty, AISHELL3-thirty and Generate-multi which is the generated dataset.]

資料集	音檔數量	總時長（小時
VCTK-thirty	11, 654	22.5
AISHELL3-thirty	13, 708	19
Generate-multi	25, 362	15.6

4. 實驗結果與分析 (Results and Analysis)

本實驗採用平均意見分數（Mean Opinion Score, MOS) 作為評估機制，分數區間為 0（低）～5（高），針對語音的整體品質進行評分，包含了流暢度、人聲相似度和有無雜訊等。隨機選取各實驗所需要的文本進行合成，由我們研究室中的 11 位研究人員參與聆聽，並對各合成語音進行評分，最後將所有分數平均做為結果。

4.1 生成資料集的品質 (Quality of the Generated Dataset)

對 3.1 生成資料集 Generated-multi 與原始資料集 AISHELL3-thirty 和 VCTK-thirty 進行比較，以確保此生成資料集的品質，由表 4 所示，可得生成資料集的分數皆在 4 分以上，表示使用生成的方式依然可獲得不錯的聲音訊號，以此資料集訓練合成器是可行的。

表 4. 資料集的 MOS。基於 VCTK-thirty 和 AISHELL3-thirty 的文本做為生成 Generatedmulti 時的文本。比較生成資料集的語音品質，生成的資料集分數在 4 分以上。

[Table 4. The MOS of the dataset. The text of the Generated dataset is based on the text of VCTK-thirty and AISHELL3-thirty. The MOS score of generated dataset is higher than 4.]

資料集	**MOS**	
	英文	中文
VCTK-thirty	4.46 ± 0.22	-
AISHELL3-thirty	-	4.73 ± 0.17
Generated-multi	4.28 ± 0.31	4.09 ± 0.62

4.2 資料前處理結果 (Results of Data Preprocessing)

由於兩種語言是分開進行資料前處理，因此在 MOS 評分，將中、英文前處理的效果分開進行比較。中文文本在訓練時皆轉為漢語拼音，於是用於評分的文本皆有經過文字轉拼音，以便系統合成，由表 5 可知，文本進行中文斷詞後，MOS 有些微的增加，另外，進行評估數字正規化的文本，為突顯正規化的效果，文本皆選用含有阿拉伯數字的中文句子，正規化後 MOS 分數由 4.02 提高到了 4.45，分數大幅的提升了，由此可知，數字正規化對於文本的重要。在英文結果的部份，選用在英文句中含有連續大寫的文本，用以評估處理頭字語的效果，然而在加入頭字語處理後，MOS 分數由 3.69 增加至 3.99，由結果可知透過前處理能提升合成之品質。

表 5. 有無進行前處理的 MOS。比較有無前處理的語音訊號品質，處理後的品質，皆有所提升。

[Table 5. The MOS of the speech which is with preprocessing or not. The quality of the speech is better after processing the text.]

語言	資料前處理流程	**w/o**	**MOS**
中	中文斷詞	w/o	4.50 ± 0.11
		w/	4.52 ± 0.18
	數字正規化	w/o	4.02 ± 0.40
		w/	4.45 ± 0.20
英	頭字語處理	w/o	3.69 ± 0.20
		w/	3.99 ± 0.15

5. 結論 (Conclusions)

我們建立的中英文語碼轉換語音合成系統，其有相當不錯的表現，透過中、英文的資料前處理大幅提升語音的品質，尤其是中文的數字正規化與英文的頭字語處理，分別由 4.02 上升至 4.45，及 3.69 至 3.99，不過整體系統依舊有進步的空間，因此，未來也將持續改進語碼轉換中，中英文的語音流暢度，以及以建立一個可分離語者資訊，單純學習文本資訊的編碼器為目標，無需再使用生成模型生成的資料集進行訓練，依然可合成多語言語碼轉換的句子。

參考文獻 (References)

Bai, Y., Tao, J., Yi, J., Wen, Z., & Fan, C. (2018). Clmad: A chinese language model adaptation dataset. In *Proceedings of 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)*, 275-279. https://doi.org/10.1109/ISCSLP.2018.8706600

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. arXiv preprint arXiv:2005.08100

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networksfor efficient and high fidelity speech synthesis. In *Proceedings of Advances in Neural Information Processing Systems*, 33, 17022-17033.

Ma, W.-Y. & Chen, K.-J. (2004). Design of ckip chinese word segmentation system. *International Journal of Asian Language Processing*, *14*(3),235-249.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558

Shi, Y., Bu, H., Xu, X., Zhang, S., & Li, M. (2020). Aishell-3: A multi-speaker mandarin tts corpus and the baselines. arXiv preprint arXiv:2010.11567

J Sun. 2012. Jieba chinese word segmentation tool.

Wang, Y.-W. (2021). Integrating hidden speaker and style information to multi-lingual and codeswitching speech synthesis. (Master's thesis). Retrieved from https://hdl.handle.net/11296/du785x

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. arXiv preprint arXiv:1804.00015.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645..