MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

End-to-End Multilingual Multi-Speaker Speech Recognition

Seki, H.; Hori, T.; Watanabe, S.; Le Roux, J.; Hershey, J.

TR2019-101 September 18, 2019

Abstract

The expressive power of end-to-end automatic speech recognition (ASR) systems enables direct estimation of a character or word label sequence from a sequence of acoustic features. Direct optimization of the whole system is advantageous because it not only eliminates the internal linkage necessary for hybrid systems, but also extends the scope of potential applications by training the model for various objectives. In this paper, we tackle the challenging task of multilingual multispeaker ASR using such an all-in-one end-to-end system. Several multilingual ASR systems were recently proposed based on a monolithic neural network architecture without languagedependent modules, showing that modeling of multiple languages is well within the capabilities of an end-to-end framework. There has also been growing interest in multi-speaker speech recognition, which enables generation of multiple label sequences from single-channel mixed speech. In particular, a multi-speaker end-to-end ASR system that can directly model one-to-many mappings without additional auxiliary clues was recently proposed. The proposed model, which integrates the capabilities of these two systems, is evaluated using mixtures of two speakers generated by using 10 languages, including codeswitching utterances.

Interspeech

End-to-End Multilingual Multi-Speaker Speech Recognition

Hiroshi Seki[1,2], *Takaaki Hori*[1], *Shinji Watanabe*[3], *Jonathan Le Roux*[1], *John R. Hershey*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), USA
[2]Toyohashi University of Technology, Japan
[3]Johns Hopkins University, USA

h123346@edu.tut.ac.jp, thori@merl.com, shinjiw@jhu.edu, leroux@merl.com

Abstract

The expressive power of end-to-end automatic speech recognition (ASR) systems enables direct estimation of a character or word label sequence from a sequence of acoustic features. Direct optimization of the whole system is advantageous because it not only eliminates the internal linkage necessary for hybrid systems, but also extends the scope of potential applications by training the model for various objectives. In this paper, we tackle the challenging task of multilingual multi-speaker ASR using such an all-in-one end-to-end system. Several multilingual ASR systems were recently proposed based on a monolithic neural network architecture without language-dependent modules, showing that modeling of multiple languages is well within the capabilities of an end-to-end framework. There has also been growing interest in multi-speaker speech recognition, which enables generation of multiple label sequences from single-channel mixed speech. In particular, a multi-speaker end-to-end ASR system that can directly model one-to-many mappings without additional auxiliary clues was recently proposed. The proposed model, which integrates the capabilities of these two systems, is evaluated using mixtures of two speakers generated by using 10 languages, including code-switching utterances.

Index Terms: end-to-end ASR, multilingual ASR, multi-speaker ASR, code-switching, encoder-decoder, CTC

1. Introduction

The expressive power of end-to-end automatic speech recognition (ASR) systems enables direct conversion from input speech feature sequences to output label sequences without any explicit intermediate representations and hand-crafted modules [1–5]. In addition to eliminating these intermediate linkage components found in hybrid systems, direct optimization of the whole system allows models to be more easily applied to different scenarios simply by changing training data and objectives.

Multilingual speech recognition is one such scenario, in which the goal is to support recognition of multiple languages. A particularly challenging case is that of *code-switching*, where speakers of multiple languages naturally switch language between or during utterances. Conventional approaches require language dependent modules and rely on a pipeline processing consisting of language identification followed by recognition of speech with a matched language-dependent system. However, recent studies have demonstrated end-to-end systems that can recognize multiple languages without language-dependent modules [6–8]. These methods eliminate the need for a language identification module, making it easier for application developers to produce systems for an arbitrary set of languages.

Whereas conventional ASR systems support recognition of speech by a single speaker, it is typically difficult or impossible to use them in scenarios where multiple people are talking simultaneously. There has recently been growing interest in dealing with such situations, with many developments in the field of single-channel multi-speaker ASR [9–17]. The goal of single-channel multi-speaker speech recognition is to recognize the speech of multiple speakers given the single-channel mixture of their acoustic signals, in a one-to-many transformation. Promising techniques have been proposed for this task, but many earlier works have required the availability of additional training information such as the isolated source signals of each speaker [11] or the phonetic state alignments [12, 13] for effective learning. Some of these also require an explicit intermediate separation stage prior to recognition [11, 13, 16, 18]. More recently, several studies have considered an end-to-end architecture to directly generate multiple hypotheses from a speech mixture without requiring additional auxiliary training signals or separation modules [15,17]. While these systems were designed for a given number of speakers, it is possible to train them for a large enough maximum number of simultaneous speakers without significant loss of performance with less speakers [12].

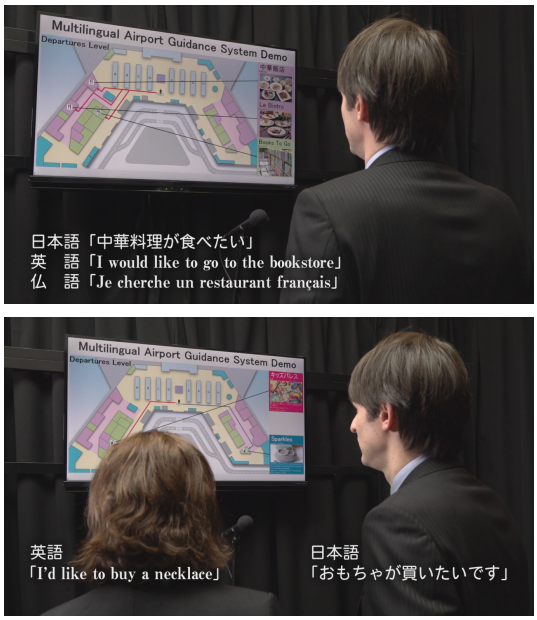

Figure 1: *Example of seamless ASR on code-switching speech (top) and multilingual multi-speaker speech (bottom).*

In this paper, we propose an unprecedented all-in-one end-to-end multilingual multi-speaker ASR system integrating end-to-end approaches for multilingual ASR and multi-speaker ASR. This system can be used to provide a seamless ASR experience, in particular improving accessibility of interfaces facing a diverse set of users. As an example of potential appli-

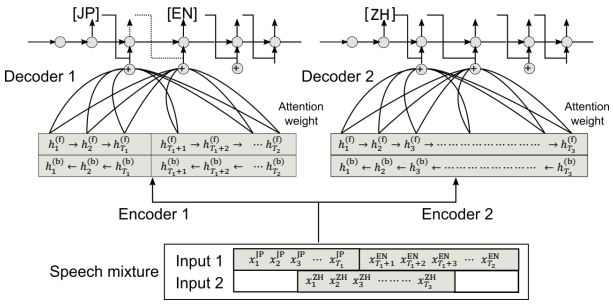

Figure 2: *Overview of the proposed end-to-end multilingual multi-speaker ASR system. The recognizer supports input in multiple languages and allows speakers to switch languages during an utterance (code-switching). For each speaker, the decoder network generates a language ID followed by a character sequence, and repeats the generation of a language ID and character sequence when the speech includes code-switching.*

cation, we developed a live demonstration of a multilingual guidance system for an airport, shown in Fig. 1. The system realizes a speech interface that can guide users to various locations within an airport. It can recognize multilingual speech with code-switching and simultaneous speech by multiple speakers in various languages without prior language settings, and provide the appropriate guidance for each query in the corresponding language. The demonstration was presented during a Japanese press event in February 2019, resulting in reports in all national TV channels in Japan as well as in Japan's top newspaper. More details and a video are available at `http://www.merl.com/demos/seamless-asr`.

Figure 2 shows an overview of the multilingual multi-speaker ASR system processing a mixture of two speakers, where one speaker first speaks in Japanese then in English, while the other speaker speaks in Chinese. In this example, the left side encoder-decoder network performs recognition of multilingual speech with code-switching between Japanese and English, and the right side network performs recognition of the Chinese part. The input speech is a mixture of two speakers and the system is expected to generate hypotheses for two character sequences, one for each speaker. Inherently, this task is a combination of existing multilingual ASR and multi-speaker ASR. In this study, we investigate whether this challenging task can be accomplished using a monolithic neural network architecture optimized through an ASR loss, the network learning to perform language-independent source separation implicitly.

2. Multilingual speech recognition

2.1. End-to-end ASR network

We employ a hybrid CTC/attention end-to-end ASR framework [19] and follow the notations of [15]. An attention-based encoder-decoder network predicts labels in a label sequence $Y = \{y_n \in \mathcal{U} | n = 1, \dots, N\}$ of length N given an input feature vector sequence O of length T and the past label history, where $\mathcal{U}$ denotes a set of character labels. At inference time, the previously emitted labels are used as past history, whereas at training time, the reference labels $R = \{r_n \in \mathcal{U} | n = 1, \dots, N\}$ are used in a *teacher-forcing* fashion. The probability of sequence Y is computed by multiplying the sequence of conditional probabilities of label y_n given the past history $y_{1:n-1}$:

$$p_{\text{att}}(Y|O) = \prod_{n=1}^{N} p_{\text{att}}(y_n | O, y_{1:n-1}). \quad (1)$$

The hybrid CTC/attention network also includes a connectionist temporal classification (CTC) sub-module computing CTC probabilities $p_{\text{ctc}}(Y|O)$ (cf. [15]). The attention network and the CTC sub-module share as input the encoder output. The CTC loss and the attention-based encoder-decoder loss are combined with an interpolation weight $\lambda \in [0, 1]$:

$$\mathcal{L}_{\text{hyb}} = \lambda \mathcal{L}_{\text{ctc}} + (1 - \lambda) \mathcal{L}_{\text{att}}, \quad (2)$$

where we define the CTC and attention losses as:

$$\mathcal{L}_{\text{ctc}} = \text{Loss}_{\text{ctc}}(Y, R) \triangleq -\log p_{\text{ctc}}(Y = R|O), \quad (3)$$

$$\mathcal{L}_{\text{att}} = \text{Loss}_{\text{att}}(Y, R) \triangleq -\log p_{\text{att}}(Y = R|O). \quad (4)$$

2.2. Augmented character set

For the recognition of multiple languages, we employ the union of all target language character sets as an augmented character set, i.e., $\mathcal{U} = \mathcal{U}^{\text{EN}} \cup \mathcal{U}^{\text{JP}} \cup \cdots$, where $\mathcal{U}^{\text{EN/JP/}\cdots}$ is a character set of a specific language, as in [6] and [8]. By using this augmented character set, likelihoods of character sequences can be computed for any language without requiring a separate language identification module. The network is trained to automatically predict the correct character sequence for the target language of each utterance.

2.3. Auxiliary language identification

Language identification symbols, such as "[EN]" and "[JP]" for English and Japanese, are further added to the augmented character set for an explicit identification of the target language and for modeling the joint distribution of a language ID and a character sequence [6, 20, 21]. The language ID is inserted at the beginning of the reference label. The final augmented character set is $\mathcal{U}^{\text{final}} = \mathcal{U} \cup \{[\text{EN}], [\text{JP}], \dots\}$.

2.4. Code-switching speech

It is natural for speakers of multiple languages to switch language between or during utterances, a phenomenon known as *code-switching*. Monolingual speakers also frequently use code-switching with foreign named entities and expressions. Code-switching speech is particularly challenging for conventional ASR systems, and typically requires combining multiple mono-lingual systems under a language identification module.

It was showed in [7] that multilingual speech recognition with code-switching can be more elegantly solved using an end-to-end multilingual ASR system trained on a dataset of code-switching speech. Because existing corpora of code-switching speech are limited and thus inadequate for large scale experiments with end-to-end frameworks, [7] instead generated a large code-switching corpus by concatenating speech from existing monolingual corpora, mostly from different speakers. Here, we use the same strategy to generate a large dataset of multi-speaker multilingual speech with code-switching, following a generation procedure described in Section 4.1.

We are aware of the limitations of such artificially generated data. In particular, our dataset does not consider acoustical nuances such as Lombard effect and appropriate room impulse responses. Several analyses also show differences between natural and artificially generated code-switching [22,23]. There are currently many efforts on the development of code-switching speech corpora [24–28], and in particular, a promising data augmentation method to increase the size of code-switching data was proposed [29]. We thus plan to consider training and evaluation of the proposed system under *real* conditions in future

works. We shall note however that the proposed system performed satisfactorily in limited testing on real code-switching speech by a single speaker as well as multilingual speech by multiple speakers in the demonstration mentioned above.

3. Multilingual multi-speaker ASR

3.1. Loss function for end-to-end multi-speaker ASR

When a speech mixture contains speech uttered by S speakers simultaneously, our encoder network generates S hidden representations from the T-frame sequence of D-dimensional input feature vectors, $O = \{o_t \in \mathbb{R}^D | t = 1, \dots, T\}$:

$$H^s = \mathrm{Encoder}^s(O),\ s = 1, \dots, S. \tag{5}$$

In the training stage, the attention-based decoder network uses reference labels $R = \{R^1, \dots, R^S\}$ for the generation of hypotheses, in a teacher-forcing fashion. There is however here an ambiguity, known as the *permutation problem* [10], as to which reference label should correspond to which estimate. Therefore, the conditional probability of the decoder network for the u-th output depends on the selected v-th reference label. The probability of the n-th label $y_n^{u,v}$ is computed by conditioning on the past reference history $r_{1:n-1}^v$:

$$p_{\text{att}}(Y_{\text{att}}^{u,v}|O) = \prod_n p_{\text{att}}(y_n^{u,v}|H^u, r_{1:n-1}^v). \tag{6}$$

During training, all possible permutations of the S sequences $R^s = \{r_1^s, \dots, r_{N_s}^s\}$ of N_s reference labels are considered, and the one leading to minimum loss is adopted for backpropagation, resulting in a permutation-free objective [10, 11, 13]. Let $\mathcal{P}$ denote the set of permutations on $\{1, \dots, S\}$. The final attention loss $\mathcal{L}_{\text{att}}$ is defined as

$$\mathcal{L}_{\text{att}} = \min_{\pi \in \mathcal{P}} \sum_{s=1}^{S} \mathrm{Loss}_{\text{att}}(Y_{\text{att}}^{s,\pi(s)}, R^{\pi(s)}), \tag{7}$$

where $\pi(s)$ is the s-th element of permutation π. As the decoder network takes more computation time than CTC, the permutation of reference labels is in practice selected based on minimizing the CTC loss only: an optimal permutation $\hat{\pi}$ is determined from the CTC network output Y_{ctc}^s (considered as a random variable) corresponding to H^s and the reference labels, as

$$\hat{\pi} = \operatorname*{argmin}_{\pi \in \mathcal{P}} \sum_{s=1}^{S} \mathrm{Loss}_{\text{ctc}}(Y_{\text{ctc}}^s, R^{\pi(s)}). \tag{8}$$

This optimal permutation is then used to compute both CTC and attention losses, which are combined as in Eq. (2):

$$\mathcal{L}_{\text{ctc}} = \sum_{s=1}^{S} \mathrm{Loss}_{\text{ctc}}(Y_{\text{ctc}}^s, R^{\hat{\pi}(s)}), \tag{9}$$

$$\mathcal{L}_{\text{att}} = \sum_{s=1}^{S} \mathrm{Loss}_{\text{att}}(Y_{\text{att}}^{s,\hat{\pi}(s)}, R^{\hat{\pi}(s)}). \tag{10}$$

By defining the augmented character set as in Section 2.3 and using a multilingual multi-speaker corpus, we can train the system to recognize simultaneous speech by multiple speakers in multiple languages.

4. Experiments

4.1. Experimental setup

A multilingual multi-speaker corpus was generated using the following corpora: WSJ (English) [30,31], CSJ (Japanese) [32], HKUST (Chinese Mandarin) [33], and Voxforge (German, Spanish, French, Italian, Dutch, Portuguese, Russian) [34] for a total of 622.7 hours and 10 languages. The generated mixtures are intended to mimic overlapped speech by two speakers, where each speaker may speak any language and change language during the utterance. Because available corpora typically do not share speakers, we here concatenate utterances in various languages uttered by different speakers. Two such streams are prepared and mixed down into a multilingual overlapped speech mixture with code-switching. We now explain this process in more detail. We first sample the number of concatenations n_{concat}^1 and n_{concat}^2 ranging from 1 to N_{concat} for code-switching within each stream. Then, n_{concat}^1 and n_{concat}^2 utterances are sampled from the union of original corpora. We limit the number of times each utterance can be selected to n_{reuse}, and prevent the same speaker from appearing in both streams to be mixed. The probability of sampling a language is proportional to the duration of its original corpus, while that of sampling an utterance within a language is uniform. Selected utterances are concatenated into respective streams, which are mixed with randomly selected SNR ranging from 0 to R dB. Since the durations of the streams to be mixed are different, we randomize the starting point of the overlapping part by padding the shorter stream with silence. These procedures are repeated until the cumulative duration d of the generated corpus reaches the total duration of the original corpora. In our experiment, N_{concat} and n_{reuse} were set to 3, and R was set to 2.5 dB.

We followed the setup of earlier work on hybrid CTC/attention-based encoder-decoder networks [3]. For the encoder network, we used the initial 6 CNN layers in the VGG network stacked with an 8-layer bi-directional long short-term memory (BLSTM) network. For the generation of multiple hypotheses, the encoder network was split at the BLSTM layer: the VGG network generates a single hidden vector, from which two speaker-differentiating 2-layer BLSTMs generate two hidden vectors. The two hidden vectors are further independently fed into the (shared) 6-layer BLSTMs and the decoder network to generate hypotheses for the utterances in the mixture. As input feature, we used 80-dimensional log mel filterbank coefficients with pitch features and their delta and delta delta features extracted using Kaldi [35]. The BLSTM layer has 320 cells in each layer and direction, and a linear projection layer with 320 units follows each BLSTM layer. The decoder network has a 1-layer LSTM with 320 cells. We used the AdaDelta algorithm [36] with gradient clipping [37] for optimization. The networks were implemented using ChainerMN [38] and optimized under synchronous data parallelism using 8 GPUs. Following the pre-training procedure in [15], we first trained a randomly initialized network using single-speaker speech without code-switching. The network was then retrained using mixed speech without code-switching, and finally using mixed speech with code-switching[1].

4.2. Results

Recognition example: Table 1 shows three examples of transcriptions generated by the proposed model. The first example contains German and Japanese utterances, where there is no code-switching. The results are almost perfect. In the second example, one stream is a concatenation of English speech followed by Chinese speech, and the other is a concatenation of two Japanese utterances. The Japanese result has a few errors

[1] [17] eliminated the pre-training step and directly trained a whole network as a mono-lingual multi-speaker ASR system.

Table 1: *Examples of multilingual multi-speaker recognition result. CER of hypotheses HYP1 and HYP2 are shown in parentheses, respectively. Errors are emphasized in red. "*" is a special token inserted to pad deletion and insertion errors for better readibility.*

Example 1	REF1:	[DE] eine höhere geschwindigkeit ist möglich
	HYP1:	[DE] eine höh*re geschwindigkeit ist möglich (CER=2.6%)
	REF2:	[JP] まずなぜこの内容を選んだかと言うと
	HYP2:	[JP] まずなぜこの内容を選んだかと言うと (CER=0.0%)
Example 2	REF1:	[EN] grains and soybeans most corn and wheat futures prices were stronger [ZH] 也是的
	HYP1:	[EN] grains and soybeans most corn and wheat futures prices were strong*k [ZH] 也是的 (CER=2.8%)
	REF2:	[JP] えーここで注目すべき**点は例十十一の二重下線部に示すように [JP] アニメですとか
	HYP2:	[JP] えーここで注目すべきい点は零十十一の二十下線部に示すように [JP] アニメですとか (CER=8.6%)
Example 3	REF1:	[EN] he noted that last week's one hundred eight point dro**p in* **the dow j*on***es industrial average resulted from a slightly weaker dollar [ZH] 吪****子其也蛮普通的 [DE] ich darf nicht
	HYP1:	[EN] it arter th*e last week's one hundred eight point cround and with* daw jumn the***defter_almove***s resulted from a fl**atly reaker dollar [ZH] 吪是想的其也蛮不同的 [DE] ich darf nicht (CER= 29.2)
	REF2:	[ES] sorte*ando los p*r*omontorios de los respaldos los golfos y penín*s *ulas formados por las **rodillas [JP] えー次の一手についてま**さまざまな******議論をしなければいけないというような状況になっています それでまーえーこれからそれぞれの研究の関連とまーえーこのセッションの****見所聞き**所というのを説明したいんですけれども [ZH] 噢
	HYP2:	[ES] sortenando los para_el turios de las respa*dos los golfos * penens bulas formados por la* car*ei*das [JP] えー次の一手についてまーさまざまなんていうのをしなければいけないとい****うな状況になっています そいでまーえーこれからそれぞれの研究の関連とまーえーこのセッションのみところききとこというのを説明したいんですけれとも [ZH] 哦 (CER=20.2)

Table 2: *Character error rates (CER) [%] of* **mixed speech** *recognized by the baseline multilingual* **single-speaker** *system.*

		# concat. utt. in softer stream			
		1	2	3	Avg.
# concat. utt. in louder stream	1	107.2	107.3	109.6	108.0
	2	107.5	100.5	102.0	103.3
	3	109.1	101.1	98.1	102.7
	Avg.	107.9	103.0	103.2	104.7

Table 3: *CERs [%] of* **mixed speech** *recognized by our proposed multilingual* **multi-speaker** *ASR system.*

		# concat. utt. in softer stream			
		1	2	3	Avg.
# concat. utt. in louder stream	1	42.9	42.0	40.3	41.7
	2	41.6	46.7	47.5	45.3
	3	40.6	47.9	50.8	46.4
	Avg.	41.7	45.5	46.2	44.5

Table 4: *Oracle CERs [%] of* **isolated speech** *for each of the utterances appearing in the mixtures used in Tables 2 and 3, recognized by the baseline multilingual* **single-speaker** *system.*

		# concat. utt. in softer stream			
		1	2	3	Avg.
# concat. utt. in louder stream	1	24.4	25.3	25.3	25.0
	2	25.2	26.1	25.9	25.7
	3	25.5	25.4	25.9	25.6
	Avg.	25.1	25.6	25.7	25.4

but these errors are in fact mostly correct in terms of pronunciation. The third example includes more complex utterances with code-switching, where each stream contains three concatenated utterances. This is the most difficult condition and the CERs are higher than in the other cases. The network did make substitution, insertion, and deletion errors, but there is no swapping of words between sentences, and language IDs are correctly estimated.

Recognition performance: Table 2 shows character error rates (CERs) for the generated multilingual multi-speaker speech recognized by the baseline multilingual single-speaker model. Results are reported separately according to the number of concatenated utterances in each stream within the mixture. We can see that the baseline model has high CERs, over 100%, because the model was trained as a multilingual single-speaker ASR system. For the evaluation of the baseline system, the generated hypothesis is duplicated to match the number of references.

Table 5: *Language ID error rates (LER) [%] of the baseline, proposed, and oracle systems.*

baseline	86.8
proposed	18.6
oracle	2.8

Table 3 shows the CERs for the generated speech recognized with the proposed model. Our model significantly reduced the CERs from the baseline model, obtaining an average CER of 44.5%, a 57.5% relative reduction from the baseline.

To investigate the lower bound of CER for the generated corpus, we evaluated the performance of the multilingual single-speaker ASR system of [7] on each of the multilingual streams used in the generated corpus, prior to mixing. This can be considered an oracle result with perfect speech separation. Table 4 shows the oracle CERs. The average CER of the oracle result was 25.4%, showing that there is still room for further performance improvement.

Language identification: Table 5 shows language identification error rates (LERs) of the baseline single-speaker system, our proposed system, and the oracle system described above. The LER was calculated by computing the edit distance between the predicted language IDs and corresponding reference language IDs. Similar to the CERs, there is a gap between the proposed and oracle results, but the obtained LERs were much better than with the baseline single-speaker ASR system.

5. Conclusion

We proposed an end-to-end multilingual multi-speaker ASR system by integrating a multilingual ASR system and a multi-speaker ASR system. The model is able to convert a speech mixture to multiple hypotheses directly without explicit separation. We evaluated the proposed model using speech mixtures involving two simultaneous speech streams in which the language can switch between 10 languages during the utterance. Our all-in-one multilingual multi-speaker system obtained 57.5% relative improvement in CER over the baseline multilingual single-speaker ASR system, and showed strong potential towards this challenging task. In future works, we plan to consider training and evaluation of the proposed system under more realistic conditions in terms of code-switching. In the same way as sequence-to-sequence models eliminated hand-crafted modules such as lexicons, we believe that removal of language-dependency and single-speaker assumption is part of the future direction towards simplicity.

6. References

[1] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[2] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[3] T. Hori, S. Watanabe, Y. Zhang, and C. William, "Advances in joint CTC-Attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. Interspeech*, Aug. 2017, pp. 949–953.

[4] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

[5] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[6] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2017.

[7] H. Seki, S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[8] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and R. Kanishka, "Multilingual speech recognition with a single end-to-end model," *arXiv preprint arXiv:1711.01694*, 2018.

[9] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech & Language*, vol. 24, no. 1, pp. 45–66, 2010.

[10] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35.

[11] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, Sep. 2016.

[12] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *arXiv preprint arXiv:1707.06527*, 2017.

[13] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 241–245.

[14] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 1, pp. 184–196, 2018.

[15] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2018.

[16] X. Chang, Y. Qian, and D. Yu, "Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks," in *Proc. Interspeech*, Sep. 2018, pp. 1586–1590.

[17] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," *arXiv preprint arXiv:1811.02062*, 2018.

[18] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[19] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[20] 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.

[21] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4749–4753.

[22] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "An analysis of a Mandarin-English code-switching speech corpus: SEAME," *Age*, vol. 21, pp. 25–8, 2010.

[23] V. Soto, N. Cestero, and J. Hirschberg, "The role of cognate words, POS tags, and entrainment in code-switching," *Proc. Interspeech*, pp. 1938–1942, Sep. 2018.

[24] "Bangor Miami corpus," http://bangortalk.org.uk/speakers.php?c=miami.

[25] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "SEAME: A Mandarin-English code-switching speech corpus in South-East asia," in *Proc. Interspeech*, 2010.

[26] A. Dey and P. Fung, "A Hindi-English code-switching corpus," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 2410–2413.

[27] E. Yilmaz, M. Andringa, S. Kingma, J. Dijkstra, F. Kuip, H. Velde, F. Kampstra, J. Algra, H. Heuvel, and D. A. van Leeuwen, "A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2016.

[28] T. Niesler *et al.*, "A first South African corpus of multilingual code-switched soap opera speech," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2018.

[29] E. Yılmaz, H. v. d. Heuvel, and D. A. van Leeuwen, "Acoustic and textual data augmentation for improved ASR of code-switching speech," *arXiv preprint arXiv:1807.10945*, 2018.

[30] "CSR-II (WSJ1) complete," vol. LDC94S13A. Philadelphia: Linguistic Data Consortium, 1994.

[31] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," vol. LDC93S6A. Philadelphia: Linguistic Data Consortium, 2007.

[32] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, vol. 2, 2000, pp. 947–952.

[33] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "HKUST/MTS: A very large scale mandarin telephone speech corpus," in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2006, pp. 724–735.

[34] "Voxforge," http://www.voxforge.org/.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2011.

[36] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[37] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks." in *Proc. International Conference on Machine learning (ICML)*, 2013, pp. 1310–1318.

[38] T. Akiba, K. Fukuda, and S. Suzuki, "ChainerMN: Scalable Distributed Deep Learning Framework," in *Proc. NIPS Workshop on ML Systems*, 2017.