



Streaming End-to-End Multilingual Speech Recognition with Joint Language Identification

C. Zhang, B. Li, T.N. Sainath, T. Strohmaier, S. Mavandadi, S. Chang, P. Haghani

Google LLC, USA

{chaoz, boboli, tsainath, strohman, sepand, shuoyiin, parisah}@google.com

Abstract

Language identification is critical for many downstream tasks in automatic speech recognition (ASR), and is beneficial to integrate into multilingual end-to-end ASR as an additional task. In this paper, we propose to modify the structure of the cascaded-encoder-based recurrent neural network transducer (RNN-T) model by integrating a per-frame language identifier (LID) predictor. RNN-T with cascaded encoders can achieve streaming ASR with low latency using first-pass decoding with no right-context, and achieve lower word error rates (WERs) using second-pass decoding with longer right-context. By leveraging such differences in the right-contexts and a streaming implementation of statistics pooling, the proposed method can achieve accurate streaming LID prediction with little extra test-time cost. Experimental results on a voice search dataset with 9 language locales shows that the proposed method achieves an average of 96.2% LID prediction accuracy and the same second-pass WER as that obtained by including oracle LID in the input. **Index Terms:** voice search, multilingual ASR, language identification, RNN-T, cascaded encoders

1. Introduction

With the rapid development of deep learning, ASR performance has significantly improved and has been applied to an increasing number of products used by many people. Although most ASR studies focus on only one language, having mixed languages in an utterance or multiple languages in a dialogue is very common in real-world applications, since approximately 43% of people in the world are bilingual with a further 17% being multilingual [1]. Hence it is required in practice to develop truly multilingual ASR systems that can handle a list of languages without knowing the language identity of a test utterance. Furthermore, in downstream tasks, such as code-switching and speech translation, both text transcriptions and their language identifiers (LIDs) are often requested together. It is beneficial to produce them together using a single multitask model built in a fully end-to-end (E2E) way.

Although E2E multilingual ASR [2–4] and language identification [5–14] can be studied separately, there exists a large number of previous work on using LID to improve multilingual E2E ASR [15–35]. One body of work shows that using oracle LID information helps multilingual models [15–20]. Our work differs from this as we predict the LIDs instead of using the oracle ones. Another body of work looks at techniques to predict LID and use the predictions in the ASR system or downstream tasks [21–29]. Much of this work focuses on LID predictions in a non-streaming system [23–26, 30–32], which does not fit into our streaming ASR setup that is important due to production constraints. Our work also differs from existing streaming work [27–29], as we use an RNN-T cascaded encoder model with a 0.9-second delay in the second-pass, making it suitable

for accurate LID predictions. Furthermore, our LID predictor is lightweight as it uses non-parametric streaming statistics pooling and increases parameters by only 0.5%. Finally, compared to previous works, our experiments were conducted on a larger-scale of 9 language locales [27–29].

In this paper, we work towards building a truly multilingual ASR system without using any oracle LID information during the test, which requires to have not only similar speech recognition accuracy as the systems with oracle LID inputs, but also the additional ability to predict LIDs in a streaming way. To achieve these goals, we propose to include a LID predictor into an E2E multilingual ASR based on RNN-T with cascaded encoders [36], which is to leverage its advantage of having two separate decoding passes with different right-contexts. To reduce the latency of showing initial ASR outputs to the users, the first-pass decoding (1st-pass) of the cascaded encoders has zero right-context, while the second-pass decoding (2nd-pass) has a 0.9-second right-context that can improve the quality of the final ASR outputs. Accurate language identification can be achieved by leveraging the 0.9-second right-context, which does not increase the latency of the 2nd-pass. The LID predictor is frame-synchronous and streaming, and the predictions can therefore be included in the input features of either the 2nd-pass decoder or the right-context encoder of the ASR. Experimental results showed that our proposed methods achieved not only accurate LID predictions but reduced WERs as well.

In the rest paper, Sec. 2 reviews RNN-T with cascaded encoders. Sec. 3 presents our proposed methods. Secs. 4 and 5 are the experimental setup and results. We conclude in Sec. 6.

2. RNN-T with Cascaded Encoders

First, we will describe the baseline RNN-T cascaded encoder system, which we build our LID prediction on top of. Given an input utterance $\mathbf{x}_{1:T}$ with T frames¹ and an output sequence $\mathbf{y} = y_{1:U}$ with U subword tokens, RNN-T is an E2E ASR that models $P(\mathbf{y}|\mathbf{x}_{1:T})$ with a single model [38]. Regarding a set of subword tokens \mathcal{V} , each target in the output layer is either a token in \mathcal{V} or \emptyset , where \emptyset is the blank symbol. Once a non-blank token y_u is emitted, the prediction network of the decoder, usually a long short-term memory (LSTM) model [38], is forwarded to derive a text representation based on $y_{1:u}$, which is fused with the acoustic representation derived from the encoder for current frame t , and the resulted vector is fed into the output layer to generate the next output symbol. Since one frame would be consumed if \emptyset is generated, the RNN-T decoder is *frame-synchronous*. In the training, let $\hat{\mathbf{y}} = \hat{y}_{1:T+U}$ be an alignment of \mathbf{y} that can be converted into \mathbf{y} by removing all occurrences of \emptyset , $\mathcal{A}(\mathbf{x}_{1:T}, \mathbf{y})$ be the reference lattice including

¹The term “frame” is used in this paper to stand for a decoding time step, which consumes 6 acoustic feature frames [37].

all possible alignments between \mathbf{y} and $\mathbf{x}_{1:T}$. The RNN-T loss is $\mathcal{L}^{\text{mnt}} = \ln P(\mathbf{y}|\mathbf{x}_{1:T})$, where $P(\mathbf{y}|\mathbf{x}_{1:T})$ is

$$P(\mathbf{y}|\mathbf{x}_{1:T}) = \sum_{\hat{\mathbf{y}} \in \mathcal{A}(\mathbf{x}_{1:T}, \mathbf{y})} \prod_{i=1}^{T+U} P(\hat{y}_i | \mathbf{x}_{1:t_i}, y_{1:u_i}),$$

and t_i and u_i are the values of t and u relevant to \hat{y}_i in $\hat{\mathbf{y}}$.

Due to production needs, we consider only streaming RNN-T, which continuously generates output texts while receiving the audio stream. That is, if a fixed-length right-context with $r \geq 0$ frames is allowed, only a partial utterance $\mathbf{x}_{1:t+r}$ can be used when calculating $\mathbf{h}_t^{\text{enc}}$. Normally, if r increases, the WER reduces; if r decreases, the latency of emitting a new word reduces. To trade-off between WER and latency, [36] proposed an improved RNN-T structure with two cascaded encoders for two separate decoding passes. In the 1st-pass, the causal encoder output $\mathbf{h}_t^{\text{enc}1}$ is derived based on $\mathbf{x}_{1:t}$ with zero latency from the right-context. In the 2nd-pass, a right-context encoder is used, whose output $\mathbf{h}_t^{\text{enc}2}$ is generated based on $\mathbf{h}_{1:t+r}^{\text{enc}1}$ (and thus $\mathbf{x}_{1:t+r}$). The 2nd-pass decoder achieves lower WERs and higher latency than the 1st-pass decoder with the r frames of extra information. Different from [36] that uses the same decoder for both decodings passes, the two decoders are separate in this paper, which makes the training loss $\mathcal{L}^{\text{casc}}$ as

$$\mathcal{L}^{\text{casc}} = \lambda \mathcal{L}^{\text{1st}} + (1 - \lambda) \mathcal{L}^{\text{2nd}}, \quad (1)$$

where \mathcal{L}^{1st} and \mathcal{L}^{2nd} are the RNN-T loss for 1st- and 2nd-pass, and $\{\lambda, 1 - \lambda\}$ are the weights on these two losses respectively.

3. Joint Language Identification and ASR

A frame-synchronous LID predictor is presented in this section which has two benefits. First, it can provide streaming LID predictions at every frame, which can easily be used by the encoder and frame-synchronous decoder of the streaming RNN-T model. Second, the long right-context of the 2nd-pass decoding of cascaded encoders is suitable for predicting LIDs, since accurate language identification often requires a few seconds of audio features that are difficult to do with the conventional low-latency streaming RNN-T. The cascaded encoder structure also allows using complementary features derived from different layers of the causal and right-context encoders to form the inputs of the LID predictor.

3.1. Joining a LID predictor with cascaded encoders

As discussed before, one of the goals of this paper is to use predicted LIDs to reduce 2nd-pass WERs. As shown in Figs. 1a and 1b, the 2nd-pass decoder and right-context encoder are both possible components to incorporate the predicted LIDs, which are both studied and compared later in Secs. 4 and 5.

In the first structure shown in Fig. 1a, the predicted LID distribution sequence $\mathbf{z}_{1:t}$ is converted to a one-hot vector sequence $\arg \max \mathbf{z}_{1:t}$, which is then concatenated with $\mathbf{h}_{1:t}^{\text{enc}2}$ and used as the input features to the 2nd-pass decoder. The input sequence of the LID predictor, $\mathbf{h}_{1:t}$, can thus be obtained by concatenating features derived from both the causal and right-context encoders, which are denoted as $\mathbf{e}_{1:t+r}^{\text{enc}1}$ and $\mathbf{e}_{1:t}^{\text{enc}2}$:

- $\mathbf{e}_{1:t+r}^{\text{enc}1}$ is extracted from an earlier causal encoder layer, which is closer to the RNN-T input layer and contains more low-level acoustic information;
- $\mathbf{e}_{1:t}^{\text{enc}2}$ is derived from a final right-context encoder layer, which has more high-level linguistic information as it is closer to the 2nd-pass decoder.

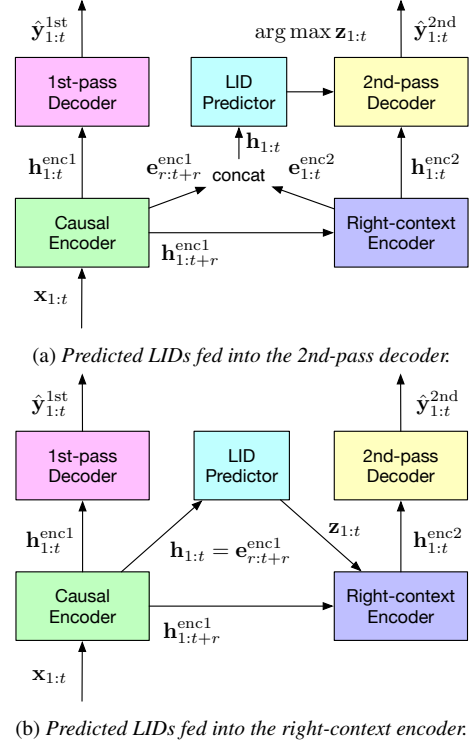


Figure 1: Two ways to incorporate the frame-synchronous LID predictor into RNN-T with cascaded encoders.

It is known that acoustic and linguistic information can be combined to improve LID prediction [9, 12, 28]. The concatenation of $\mathbf{e}_{r:t+r}^{\text{enc}1}$ and $\mathbf{e}_{1:t}^{\text{enc}2}$ allows the LID predictor to leverage such complementary information easily.

Alternatively, as shown in Fig. 1b, LID predictions are fed into the right-context encoder. That is, $\mathbf{z}_{1:t+r}$ and $\mathbf{h}_{1:t+r}^{\text{enc}1}$ are concatenated to form the input features of the right-context encoder. The input sequence to the LID predictor is $\mathbf{h}_{1:t} = \mathbf{e}_{r:t+r}^{\text{enc}1}$. In addition to the advantage that both 2nd-pass decoder and right-context encoder can benefit from having the predicted LID information, there also exist two disadvantages of this structure. First, only causal encoder features are used for LID prediction, which reduces the diversity of the inputs to the predictor. Second, due to the design of the attention matrices of the streaming Conformer encoder, the right-context allowed for the first encoder layer is often smaller than that allowed for the decoder (e.g. 3 frames vs. 15 frames in [36, 37] and this paper), which further limits the performance of LID predictor.

At last, for both structures, the final training loss $\mathcal{L}^{\text{asr+lid}}$ is obtained by extending $\mathcal{L}^{\text{casc}}$ to have an additional term for training the LID predictor as a classifier, which is

$$\mathcal{L}^{\text{asr+lid}} = \mathcal{L}^{\text{casc}} + \alpha \mathcal{L}^{\text{lid}}, \quad (2)$$

where α is a scalar weight. \mathcal{L}^{lid} is the cross-entropy loss with $\mathcal{L}^{\text{lid}} = \mathbf{l}_t \ln(\mathbf{l}_t / \mathbf{z}_t)$, where \mathbf{l}_t is the one hot LID label of t .

3.2. Statistics pooling for streaming LID predictions

LID prediction is more accurate when using longer audio features [7, 14]. To facilitate the use of all past frames but still make per-frame predictions, non-parametric statistics pooling [8] is used. This pooling converts $\mathbf{h}_{1:t}$ into the concatenation of its mean and standard deviation, $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$. $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ can be

computed in a streaming way:

$$\boldsymbol{\mu}_t = \theta(\mathbf{h}_{1:t})/t \quad (3)$$

$$\boldsymbol{\sigma}_t^2 = (\theta(\mathbf{h}_{1:t}^2) - 2\boldsymbol{\mu}_t\theta(\mathbf{h}_{1:t}) + t\boldsymbol{\mu}_t^2)/t, \quad (4)$$

where $\theta(\mathbf{h}_{1:t}) = \sum_{t'=1}^t \mathbf{h}_{t'}$ and $\theta(\mathbf{h}_{1:t}^2) = \sum_{t'=1}^t \mathbf{h}_{t'}^2$. For $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\sigma}_{t+1}$, both $\theta(\mathbf{h}_{1:t+1})$ and $\theta(\mathbf{h}_{1:t+1}^2)$ are obtained easily by adding \mathbf{h}_{t+1} and \mathbf{h}_{t+1}^2 to $\theta(\mathbf{h}_{1:t})$, $\theta(\mathbf{h}_{1:t}^2)$. After converting $\mathbf{h}_{1:t}$ into a concatenated vector $[\boldsymbol{\mu}_t; \boldsymbol{\sigma}_t]$ with the statistics pooling, another two fully connected (FC) layers followed by a softmax output layer are used to transform $[\boldsymbol{\mu}_t; \boldsymbol{\sigma}_t]$ into the LID distribution \mathbf{z}_t . As a result, our frame-synchronous LID predictor is efficient for handling streaming data that only requires a small amount computational cost during test.

4. Experimental Setup

4.1. Data

Experiments were conducted on a dataset with 9 language locales collected from Google’s Voice Search traffic, which are: US English (en-US), UK English (en-GB), French (fr-FR), Italian (it-IT), Germany (de-DE), US Spanish (es-US), ES Spanish (es-ES), Taiwan Chinese (zh-TW) and Japanese (ja-JP). The training data is mixed with all language locales. There are in total 214.2M utterances corresponding to 142.3K hours of speech data. en-US and en-GB take about 25% and 5% of the training data, while each of the rest 7 locales takes about 10% of the data. The SpecAugment method [39] is used to improve ASR robustness against noisy conditions. The testing utterances are also sampled from Google’s Voice Search traffic with a maximum duration constraint of 5.5-second long for each utterance. The test sets have no overlapping from the training set for evaluation purposes. All data are anonymised and hand-transcribed, and abide by Google AI Principles [40]. In addition, all data is normalised to have zero-mean and unit-variance based on the statistics collected from the whole training set.

4.2. Model

The acoustic features are 80-dimensional (-dim) log-Mel filter bank features computed using a 32ms frame length and a 10ms shift. Acoustic feature vectors from 3 contiguous frames are stacked and subsampled to form a 240-dim input vector with a 30ms frame rate, which is then transformed using a linear projection to 512-dim and added with positional embeddings.

Regarding the causal encoder, twelve Conformer encoder blocks with 8-head self-attention and a convolution kernel size of 15 are used to further transform the stacked features. A concatenation operation is performed after the 3rd block to achieve a time reduction rate of 2, and the resulted 1024-dim vectors are transformed by the 4th Conformer block and then projected back to 512-dim using another linear transform. Afterward comes with another 8 Conformer blocks followed by a final linear normalisation layer. These layers combined together form the causal-encoder with 110M parameters. The right context of every causal-encoder layer is 0 frame.

For the right-context encoder, a 512-dim linear projection is first used to transform its input features, followed by five 512-dim Conformer blocks and a final linear normalization layer, which altogether results in 50M parameters. The total right context of the right-context-encoder is 15 frames, which results in a 0.9-second latency of the second-pass decoder. Such 15-frame right-context is evenly distributed to the five Conformer blocks, with each of them having a 3-frame right context.

There are in total 16,384 word pieces used as the subword tokens for the 9 language locales combined and are modelled with a single softmax output layer following [4]. The two separate decoders use the same structure with 33M parameters each, which consists of two layers of 2,048-dim LSTM with a 640-dim linear projection in the prediction network. Our LID predictor uses two 512-dim FC layers that result in 0.7M–1M parameters depending on the size of its input feature sequence, which takes less than 0.5% of the total model parameters.

All models are trained in Tensorflow with the Lingvo toolkit on Google’s TPU V3 with a global batch size of 4,096 utterances. Models are optimized using synchronized stochastic gradient descent based on the Adam optimiser with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. λ and α in Eqns. (1) and (2) are set to 0.5 and 0.05. The models are tested without any external LMs.

5. Experimental Results

5.1. Summary of 800K-step ASR results

ASR results of the baseline and proposed cascaded encoder systems are listed in Table 1. S00 is the system without using any LID information. S01–S02 are those with the oracle one-hot LID vectors being used to augment the input features of different model components for both training and test. S20 and S30 are with our proposed structures shown in Fig. 1a and Fig. 1b respectively. Comparing S01 with S00, including oracle LID information throughout model reduced 1st-pass and 2nd-pass WERs both by a relative of 3%, which is the maximum WER reduction can be achieved with predicted LIDs. Compared with S01, S20 achieved the same 2nd-pass WER, showing the effectiveness of the proposed structure that feeds the predicted LID into the 2nd-pass decoder. S20 1st-pass WER is similar to S00 and worse than S01, which is expected since the predicted LIDs are not used in the 1st-pass. The estimated 2nd-pass WER gain of our proposed system S20 compared to the baseline S00 is 0.2% (computed from ~ 0.4 M utterances across the 9 locales), which is **statistically significant** at the 5% level. Furthermore, since the 2nd-pass WER is the final WER of the system, the Table shows that we can achieve the first goal of this paper, namely that the same final ASR accuracy can be obtained using predicted (S20) and oracle LIDs (S01).

The WER reduction over S00 obtained by S30 is only 33% of that achieved by S20. When reducing the right-context from 15 to 3 to meet the 3-frame right-context of the first Conformer block of the right-context encoder, the resulting S31 system WERs were slightly better than S00 and received an improvement in ASR accuracy with the predicted LIDs. These reveal that the structure in Fig. 1a is better than that in Fig. 1b. Furthermore, S20 can be viewed as replacing the oracle LIDs in S02 with the predicted ones. Since S20 had lower 2nd-pass WER than S02, not only including the predicted LIDs in the 2nd-pass decoder, but also having a shortcut for gradient back-propagation between the 2nd-pass decoder and causal encoder helped to reduce 2nd-pass WER. In the next section, we will validate this finding by analysing S20.

5.2. Input and output settings for the LID predictor

In this section, S20 is analysed to understand the impact of different input and output settings. All results in Table 2 and 3 are produced by systems trained for 200K steps, instead of the final 800K-step systems used in Table 1. Table 2 shows the results of deriving $\mathbf{e}_{1:t+r}^{\text{enc}1}$ and $\mathbf{e}_{1:t}^{\text{enc}2}$ from different encoder layers to form $\mathbf{h}_{1:t}$ in Fig. 1a systems. Deriving $\mathbf{e}_{1:t+r}^{\text{enc}1}$ from the

Table 4: 800K-step LID prediction %accuracy breakdown to individual languages/locales. “en-X” is the cluster of “en-GB” & “en-US”, and “es-X” is the cluster of “es-ES” & “es-US”. The systems used are S20 with or without statistics pooling. “Frame” is the frame index t of $\mathbf{e}_{1:t}^{\text{enc}2}$ in Fig. 1a, and “avg.” refers to the average of all possible t .

Pooling	Frame	de-DE	en-GB	en-US	en-X	es-ES	es-US	es-X	fr-FR	it-IT	ja-JP	zh-TW	Mean
w/	avg.	95.5	82.0	96.8	–	87.3	89.7	–	95.3	94.6	97.4	97.2	92.9
w/o	avg.	92.3	75.2	95.5	–	81.3	84.5	–	91.7	90.5	95.3	95.1	89.0
	avg.	95.5	–	–	97.5	–	–	96.0	95.3	94.6	97.4	97.2	96.2
	0	86.2	–	–	92.7	–	–	86.2	85.6	80.8	92.6	92.4	88.1
w/	15	93.7	–	–	96.3	–	–	94.0	93.8	91.7	96.9	96.6	94.7
	30	97.1	–	–	98.2	–	–	97.3	97.1	96.2	98.8	98.5	97.6
	∞	98.5	–	–	99.1	–	–	98.7	98.5	98.3	99.2	99.0	98.8

Table 1: 800K-step ASR results of key systems, where “1st” and “2nd” are 1st- and 2nd-pass decoding results in %WER.

ID	System	1st	2nd
S00	no LID	9.1	8.0
S01	oracle LID to causal encoder	8.9	7.8
S02	oracle LID to 2nd-pass decoder	9.1	7.9
S20	Fig. 1a system with $r=15$	9.1	7.8
S30	Fig. 1b system with $r=15$	9.1	7.9
S31	Fig. 1b system with $r=3$	9.1	7.9

4th Conformer block resulted in lower 1st-pass WERs due to the regularisation effect. By combining that $\mathbf{e}_{1:t+r}^{\text{enc}1}$ with the $\mathbf{e}_{1:t}^{\text{enc}2}$ derived from the 5th Conformer block of the right-context encoder, higher LID prediction accuracy can be achieved that leverages the complementary acoustic and linguistic information. This setting (last row of Table 2) is used in S20.

Table 2: 200K-step Fig. 1a system results using different inputs for the LID predictor, where b_n refers to outputs from the n -th Conformer block. “1st” and “2nd” are the 1st-pass and 2nd-pass decoding results in %WER; “9-cl” is the LID %accuracy averaged over the 9 language locales and across time.

$\mathbf{e}_{1:t+r}^{\text{enc}1}$	$\mathbf{e}_{1:t}^{\text{enc}2}$	1st	2nd	9-cl
b_4	–	9.9	8.5	89.8
–	b_5	10.0	8.5	90.6
b_4	b_5	9.9	8.5	91.5

Next, different ways of feeding the predicted LID distribution $\mathbf{z}_{1:t}$ into the 2nd-pass decoder are compared by training different Fig. 1a systems. Compared with feeding $\mathbf{z}_{1:t}$ directly, feeding $\arg \max \mathbf{z}_{1:t}$ as a sequence of 9-dim one-hot vectors achieved lower WERs and higher LID prediction accuracy. When feeding $\text{sg}(\arg \max \mathbf{z}_{1:t})$, where $\text{sg}(\cdot)$ means the input tensor will have zero gradients, the shortcut created by the LID predictor between the 2nd-pass decoder and causal encoder is removed. Although this improves the 1st-pass WER and LID prediction accuracy, it increases the 2nd-pass WER. This shows that some improvements in the 2nd-pass decoding when using $\arg \max \mathbf{z}_{1:t}$ come from the shortcut. Finally, clustering “en-GB” and “en-US” into “en-X” and clustering “es-ES” and “es-US” into “es-X”, the 9 language locales are merged into 7 distinct languages. By using the 7 languages instead of 9 locales in both training and test, the resulted system (last row of Table 3) has worse performance compared to $\arg \max \mathbf{z}_{1:t}$. As a result, $\arg \max \mathbf{z}_{1:t}$ is used for S20 in Table 1 to achieve a good trade-off between the 2nd-pass WER and LID prediction accuracy.

Table 3: 200K-step Fig. 1a system results with different ways of using predicted LIDs in the 2nd-pass decoder. “1st” and “2nd” are the 1st-pass and 2nd-pass decoding results in %WER. “9-cl” and “7-cl” are the LID prediction results in %accuracy avg. across time for the 9 language locales and 7 languages.

System	1st	2nd	9-cl	7-cl
$\mathbf{z}_{1:t}$	10.0	8.6	91.3	95.3
$\arg \max \mathbf{z}_{1:t}$	9.9	8.5	91.5	95.5
$\text{sg}(\arg \max \mathbf{z}_{1:t})$	9.9	8.6	91.6	95.4
$\text{cluster}(\arg \max \mathbf{z}_{1:t})$	10.0	8.6	–	95.4

5.3. Analysis of streaming LID predictions

At last, detailed LID prediction results of Fig. 1a structured systems are given in Table 4. From the results, S20 achieved a 92.9% LID classification accuracy averaged across time, which outperformed S20 without statistics pooling (2nd row in Table 4). By further evaluating the clustered 7 languages, the averaged LID prediction accuracy of S20 reached 96.2%. Evaluating LID performance at different frames in the LID prediction sequence, the LID accuracy increases when more frames are available in a streaming utterance, and the final LID accuracy based on complete utterances is 98.8%. In particular, from 0th frame to 15th frame, “de-DE”, “es-X”, “fr-FR”, and “it-IT” received more obvious increases in accuracy than “en-X”, “ja-JP” and “zh-TW”, which is perhaps due to their similarity to “en-X” that can require longer feature sequences to be well discriminated. These results show that by adding an efficient LID predictor, RNN-T with cascaded encoders can produce LID accurately in a streaming way, which fulfills our second goal.

6. Conclusions

We propose to build a truly multilingual E2E ASR system that combines ASR with LID prediction based on the structure of RNN-T with cascaded encoders. By leveraging on the long right-context in the second-pass decoding that is useful for LID prediction, a lightweight predictor is added to the ASR structure to achieve accurate frame-synchronous LID prediction in a streaming way. Experiments on a large-scale multilingual voice search setup with 9 language locales showed, by only increasing 0.5% of the number of model parameters and little test-time cost, the LID predictor achieved an average of 96.2% accuracy across time when evaluated on 7 languages clustered from the 9 locales. By feeding the predicted LID into the second-pass decoder, our multilingual ASR that does not require knowing LIDs can achieve the same final WER as the system that includes the oracle LID as extra input. It is possible to apply our method to handle code-switch with mixed language data.

7. References

- [1] iLanguages.org, “Multilingual people,” 2018. [Online]. Available: <http://ilanguages.org/bilingual.php>
- [2] J. Cho, M. Baskar, R. Li, M. Wiesner, S. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, “Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modelling,” in *Proc. SLT*, Athens, 2018.
- [3] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, “Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters,” in *Proc. Interspeech*, Shanghai, 2020.
- [4] B. Li, R. Pang, T. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. Huang, M. Ma, and J. Bai, “Scaling end-to-end models for large-scale multilingual ASR,” *arXiv:2104.14830*, 2021.
- [5] R. Fér, P. Matějka, F. Gréz, O. Plchot, and J. Černocký, “Multilingual bottleneck features for language recognition,” in *Proc. Interspeech*, Dresden, 2015.
- [6] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. Moreno, and J. Gonzalez-Rodriguez, “Frame-by-frame language identification in short utterances using deep neural networks,” *Neural Networks*, vol. 64, pp. 49–58, 2015.
- [7] J. Pešán, L. Burget, and J. Černocký, “Sequence summarizing neural networks for spoken language recognition,” in *Proc. Interspeech*, San Francisco, 2016.
- [8] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, “Spoken language recognition using x-vectors,” in *Proc. Odyssey*, Les Sables d’Olonne, 2018.
- [9] S. Wang, L. Wan, Y. Yu, and I. L. Moreno, “Signal combination for language identification,” *arXiv:1910.09687*, 2019.
- [10] L. Wan, P. Sridhar, Y. Yu, Q. Wang, and I. Lopez-Moreno, “Tuplemax loss for language identification,” in *Proc. ICASSP*, 2019.
- [11] A. Titus, J. Silovsky, N. Chen, R. Hsiao, M. Young, and A. Ghoshal, “Improving language identification for multilingual speakers,” in *Proc. ICASSP*, Barcelona, 2020.
- [12] C. Chandak, Z. Raeesy, A. Rastrow, Y. Liu, X. Huang, S. Wang, D. Joo, and R. Maas, “Streaming language identification using combination of acoustic representations and ASR hypotheses,” *arXiv:2006.00703*, 2020.
- [13] S. Punjabi, H. Arsikere, Z. Raeesy, C. Chandak, N. Bhave, A. Bansal, M. Müller, S. Murillo, A. Rastrow, S. Garimella, R. Maas, M. Hans, A. Mouchtaris, and S. Kunzmann, “Streaming end-to-end bilingual ASR systems with joint language identification,” *arXiv:2007.03900*, 2020.
- [14] A. Tjandra, D. Choudhury, F. Zhang, K. Singh, A. Conneau, A. Baevski, A. Sela, Y. Saraf, and M. Aulli, “Improved language identification through cross-lingual self-supervised learning,” *arXiv:2107.04082*, 2021.
- [15] S. Kim and M. Seltzer, “Towards language-universal end-to-end speech recognition,” in *Proc. ICASSP*, Calgary, 2018.
- [16] S. Toshniwal, T. Sainath, R. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, “Multilingual speech recognition with a single end-to-end model,” in *Proc. ICASSP*, Calgary, 2018.
- [17] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, “Leveraging language ID in multilingual end-to-end speech recognition,” in *Proc. ASRU*, Singapore, 2019.
- [18] N. Gaur, B. Farris, P. Haghani, I. Leal, P. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, “Mixture of informed experts for multilingual speech recognition,” in *Proc. ICASSP*, 2021.
- [19] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, “Transformer-transducer for code-switched speech recognition,” in *Proc. ICASSP*, Toronto, 2021.
- [20] L. Zhou, J. Li, E. Sun, and S. Liu, “A configurable multilingual model is all you need to recognize all languages,” in *Proc. ICASSP*, Singapore, 2022.
- [21] J. Weiner, N. Vu, D. Telaar, F. Metze, T. Schultz, D.-C. Lyu, E.-S. Chng, and H. Li, “Integration of language identification into a recognition system for spoken conversations containing code-switches,” in *Proc. Interspeech*, Cape Town, 2012.
- [22] J. Gonzalez-Dominguez, D. Eustis, I. Lopez-Moreno, A. Senior, F. Beaufays, and P. Moreno, “A real-time end-to-end multilingual speech recognition architecture,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9(4), pp. 749–759, 2015.
- [23] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition,” in *Proc. ICASSP*, Calgary, 2018.
- [24] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, “Towards end-to-end code-switching speech recognition,” *arXiv:1810.13091*, 2018.
- [25] Z. Zeng, Y. Khassanov, V. Pham, H. Xu, E. Chng, and H. Li, “On the end-to-end solution to Mandarin-English code-switching speech recognition,” in *Proc. Interspeech*, Graz, 2019.
- [26] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, “Investigating end-to-end speech recognition for Mandarin-English code-switching,” in *Proc. ICASSP*, Brighton, 2019.
- [27] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, “Towards code-switching ASR for end-to-end CTC models,” in *Proc. ICASSP*, Brighton, 2019.
- [28] S. Punjabi, H. Arsikere, Z. Raeesy, C. Chandak, N. Bhave, A. Bansal, M. Müller, S. Murillo, A. Rastrow, A. Stolcke, J. Droppo, S. Garimella, R. Maas, M. Hans, A. Mouchtaris, and S. Kunzmann, “Joint ASR and language identification using RNN-T: An efficient approach to dynamic language switching,” in *Proc. ICASSP*, Toronto, 2021.
- [29] V. Joshi, A. Das, E. Sun, R. Mehta, J. Li, and Y. Gong, “Multiple softmax architecture for streaming multilingual end-to-end ASR systems,” in *Proc. Interspeech*, Brno, 2021.
- [30] S. Watanabe, T. Hori, and J. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *Proc. ASRU*, Okinawa, 2017.
- [31] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. Hershey, “An end-to-end language-tracking speech recognizer for mixed-language speech,” in *Proc. ICASSP*, Calgary, 2018.
- [32] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinzaki, “Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning,” in *Proc. Interspeech*, Shanghai, 2020.
- [33] S. Zhang, J. Yi, Z. Tian, J. Tao, and Y. Bai, “RNN-transducer with language bias for end-to-end Mandarin-English code-switching speech recognition,” in *Proc. ISCSLP*, Hong Kong, 2021.
- [34] K. Hu, H. Sak, and H. Liao, “Adversarial training for multilingual acoustic modeling,” *arXiv:1906.07093*, 2019.
- [35] B. Yan, C. Zhang, M. Yu, S.-X. Zhang, S. Dalmia, D. Berrebbi, C. Weng, S. Watanabe, and D. Yu, “Joint modeling of code-switched and monolingual ASR via conditional factorization,” in *Proc. ICASSP*, Singapore, 2022.
- [36] A. Narayanan, T. Sainath, R. Pang, J. Yu, C.-C. Chiu, R. Prabhavalkar, E. Variani, and T. Strohmaier, “Cascaded encoders for unifying streaming and non-streaming ASR,” in *Proc. ICASSP*, Toronto, 2021.
- [37] Y. He, T. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. Sim, T. Bagby, S. Chang, R. Rao, and A. Gruenstein, “Streaming end-to-end speech recognition for mobile devices,” in *Proc. ICASSP*, 2019.
- [38] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, 2013.
- [39] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. Interspeech*, 2019.
- [40] Google, “Artificial Intelligence at Google: Our Principles.” [Online]. Available: <https://ai.google/principles/>