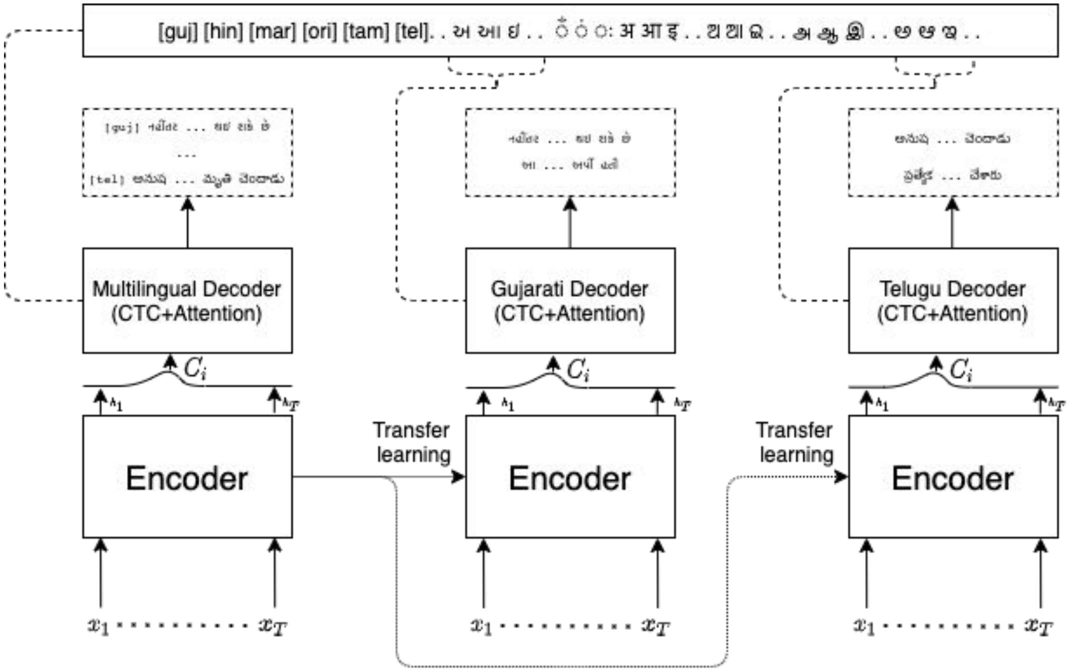
The Dialpad ASR System for the MUCS 2021 (Subtask 1)

Shreekantha Nadig, Riqiang Wang, Wang Yau Li, Jeffrey Michael, Frederic Mailhot, Simon Vandieken, Jonas Robertson

dialpad

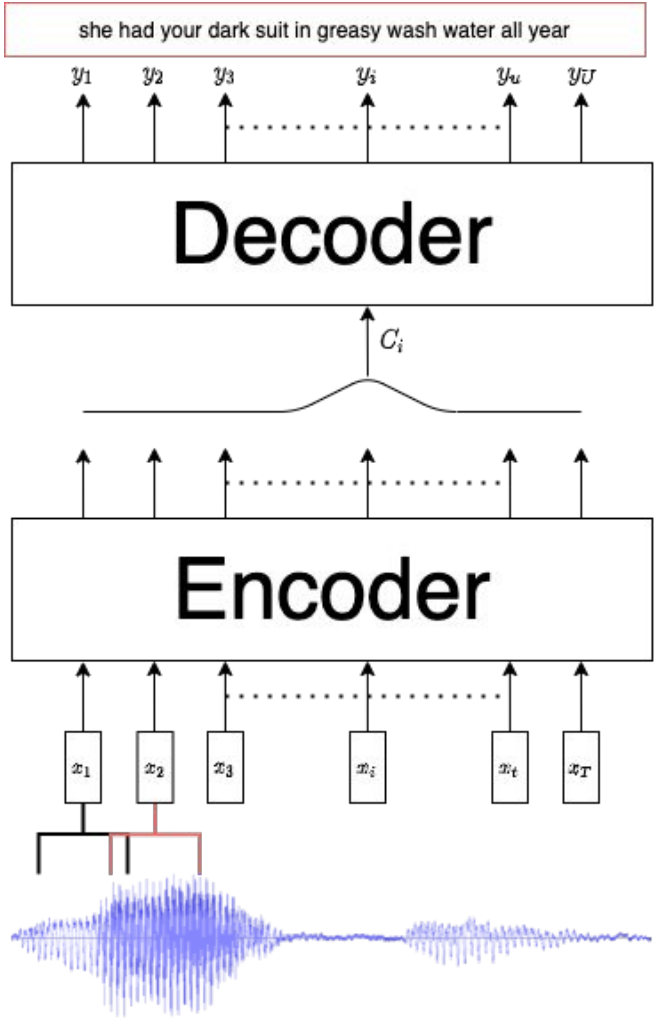


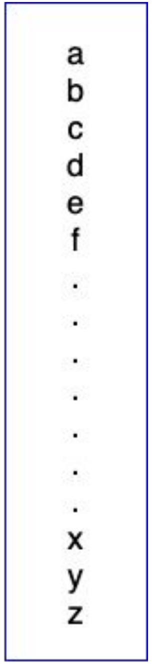
Our approach

- Combine best of published works
- UTF-8 characters as units
- Transfer learning + Fine tuning

Name	Description
B0	Baseline encoder-decoder
B1	B0's encoder + monolingual decoder
B3	B0 but with transliterated latin script
C0	B0 + explicit LID subtask
C1	B3's encoder + explicit LID decoder

dialpad

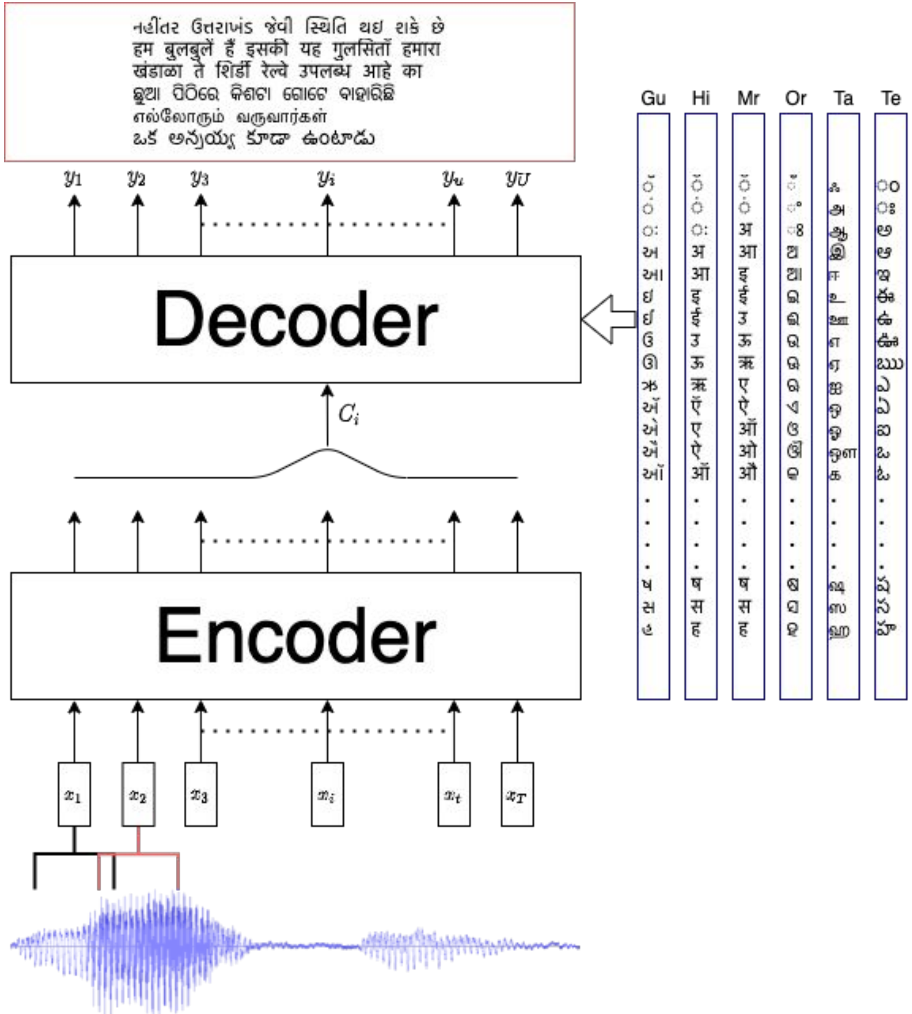




Common features

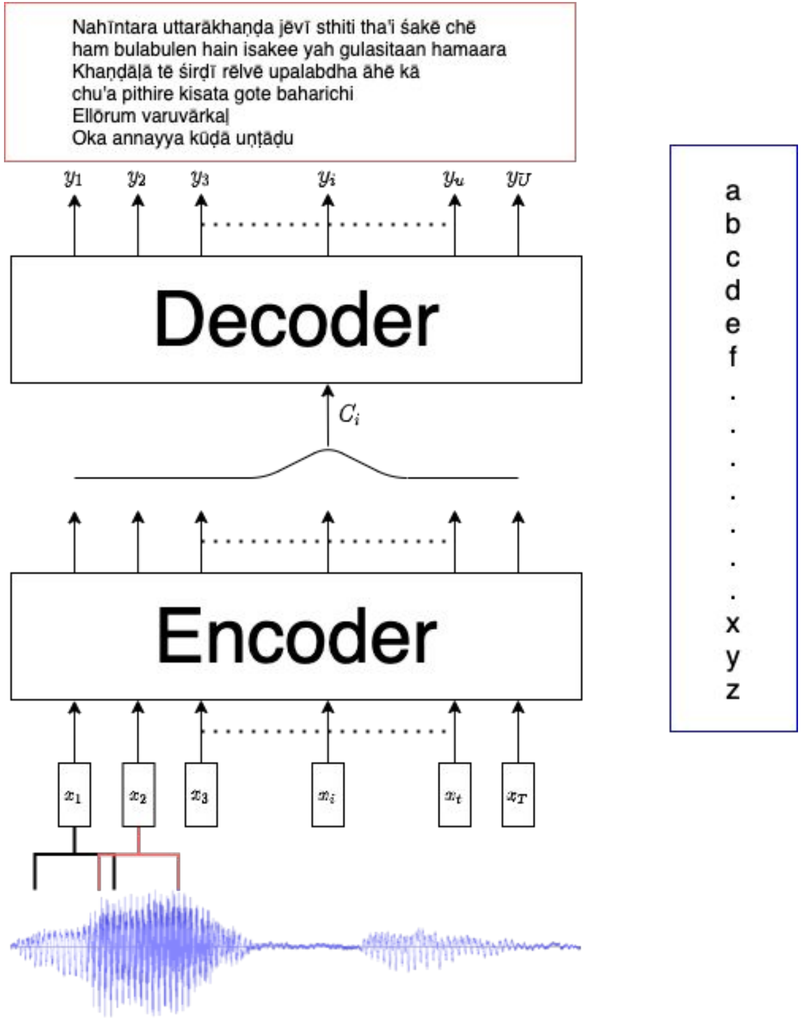
- Encoder-Decoder architecture
- CTC+ATT MTL training/decoding
- 80-dimensional log-Mel filter bank features using `torchaudio`
- ESPnet for both E2E and RNNLM
- Sentencepiece for tokenization
- RNNLMs trained with same vocabulary as that of the E2E model
- Encoder: 6 layer VGG-BLSTM with 1024 units
- Decoder: 2 layer LSTM with 1024 units
- Attention: location-aware
- No external audio data used

dialpad



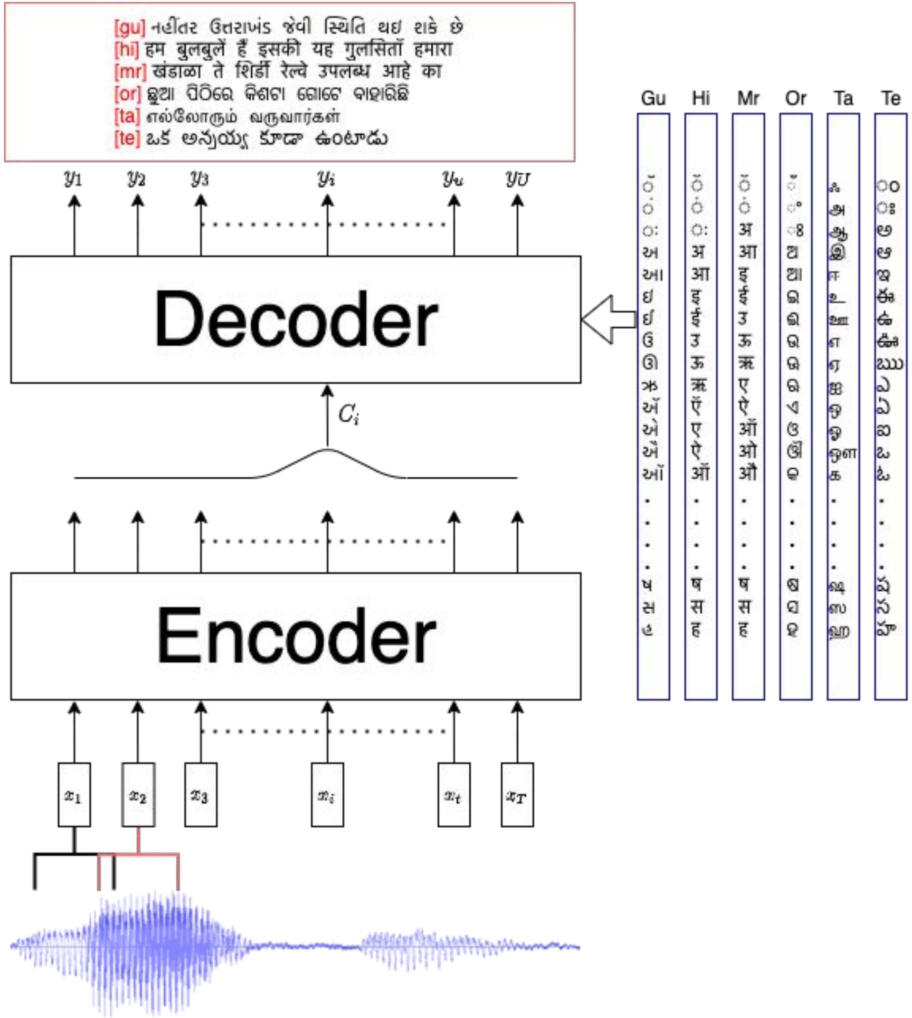
Language independent (B0)

- Proposed by Watanabe et al.
- Vocabulary of all languages are combined
- Network is trained for CTC+Attention cost functions
- Not explicitly optimized for LID
- Must perform LID implicitly
- Performs well in choosing the right orthography for the output
- No code-switching in orthography during decoding
- 99.98% accurate on the dev set



Transliteration pre-training (B3)

- Transliterate all languages to Latin alphabet using `indic-trans` (Bhat et al.)
- Hypothesis: Force the Encoder to learn representations common across languages
- Acts as a bottleneck
- Initialize the Encoder from this step to a multilingual model **(C1)**, stand-alone LID model **(L0)**



Joint LID+ASR (C0)

Introduced by Watanabe et al.

Append LID tag to the beginning of each training transcript

CTC+Attention training

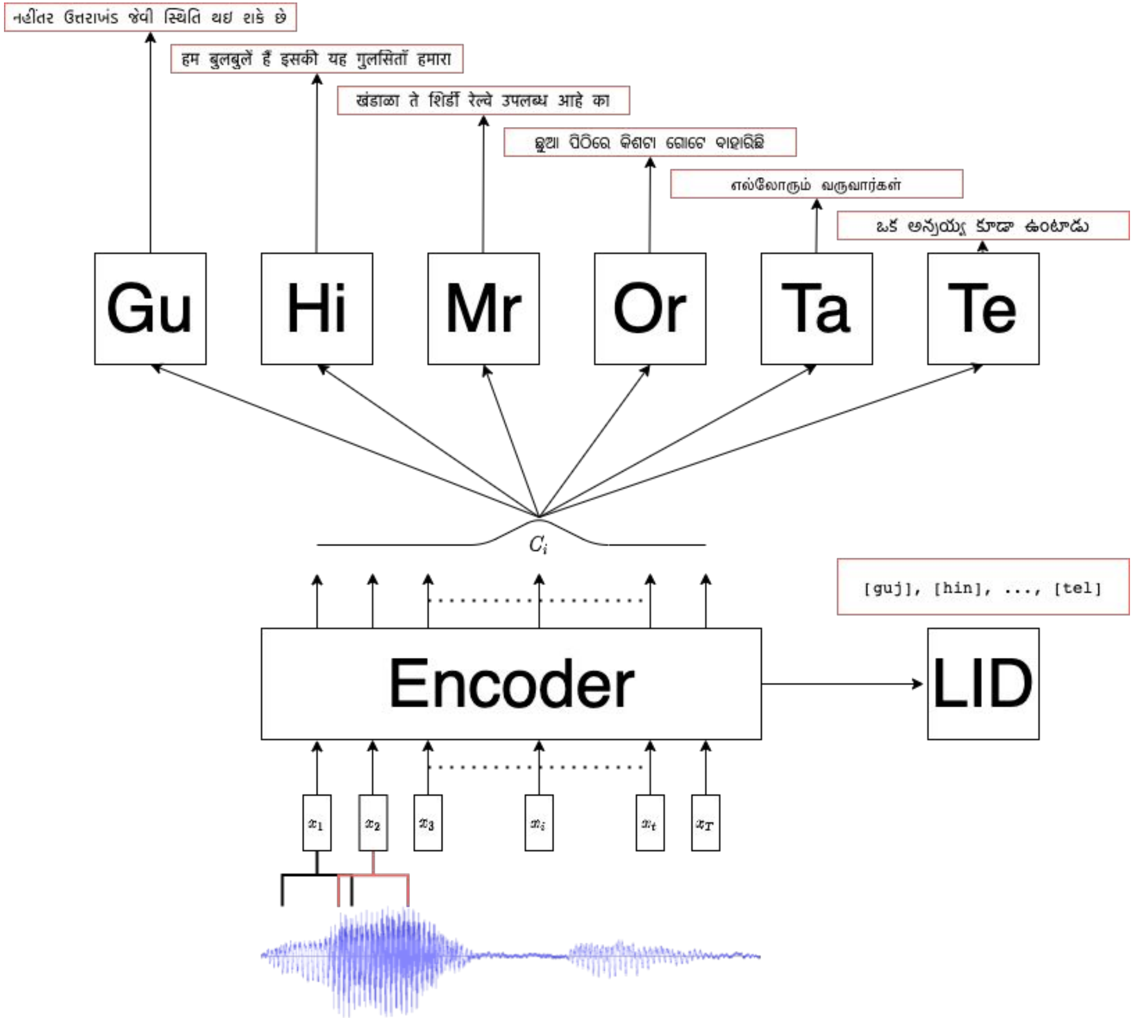
Model must first perform LID, then ASR

Achieves good LID accuracy on the dev set

CTC-only decoding: 99.93%

ATT-only decoding: 99.92%

Attends to different parts of the utterance for different languages for LID



Multi Decoder

- Introduced by Pratap et al.
- Separate decoder for each language
- Beneficial for using language-specific LMs
- Freeze Encoder and train decoders
- Un-freeze Encoder and fine-tune decoders
- Confidence based decoding approach during inference
- LID accuracy of 98% with 1-best and 99.1% with 20-best re-scoring

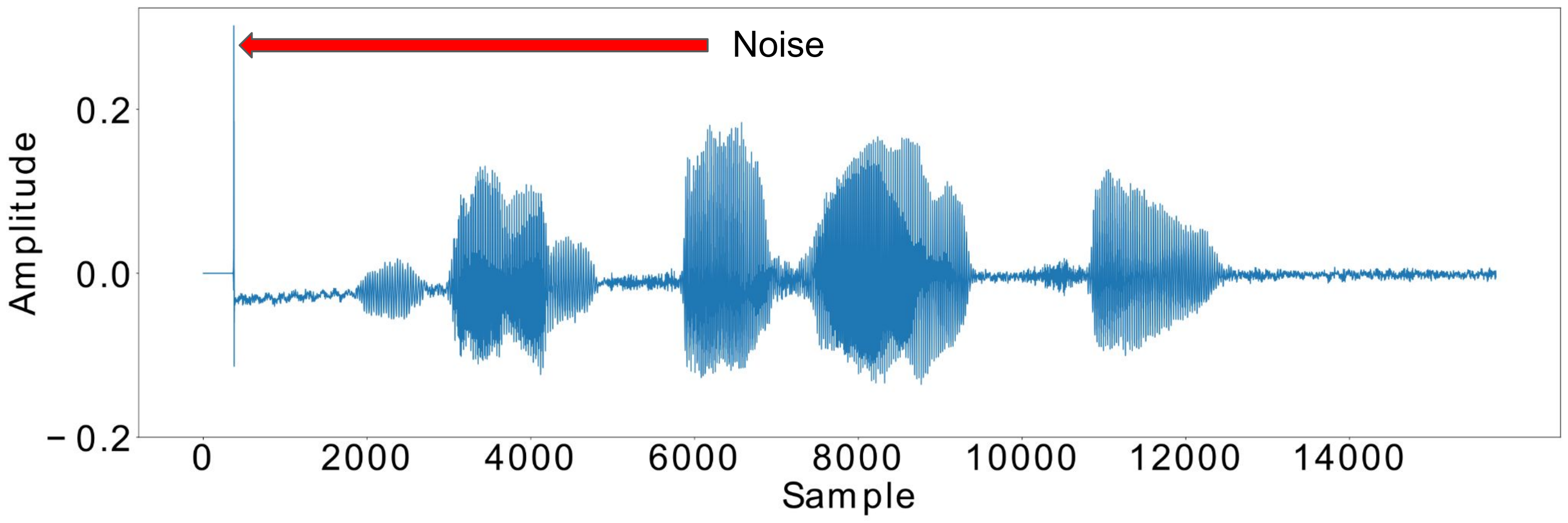
dialpad

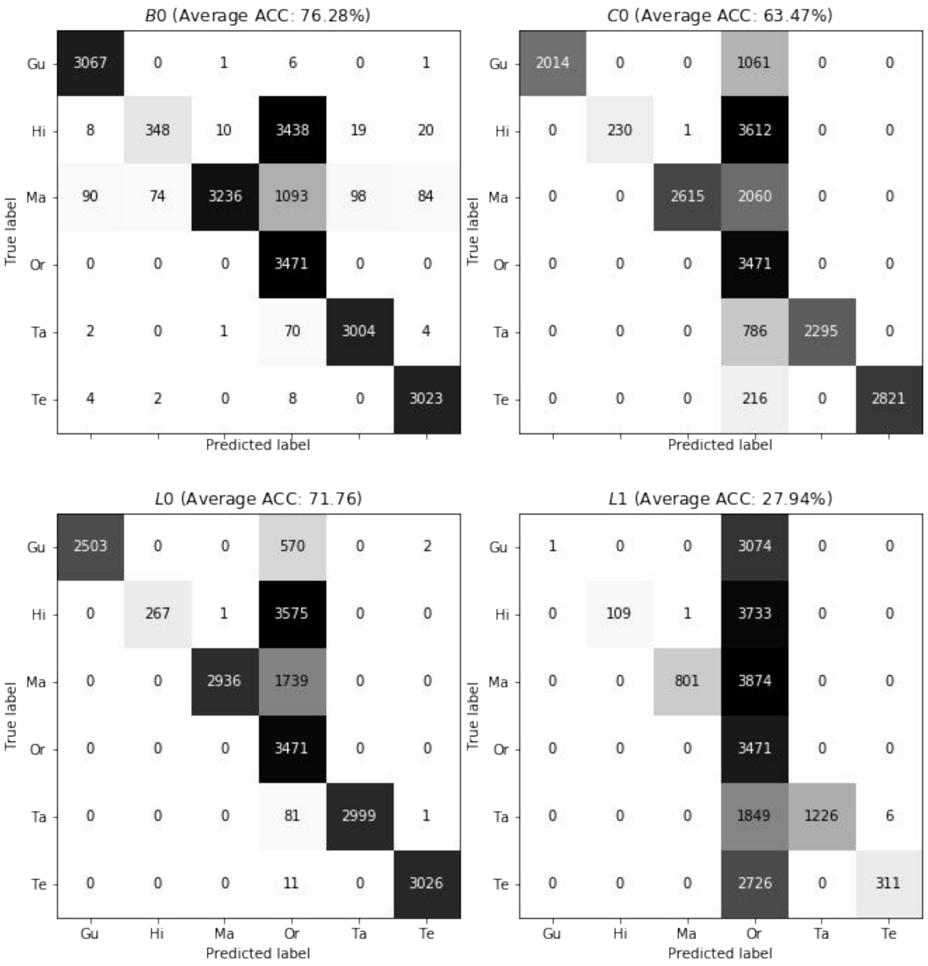
Results on the development and (held-out) test sets

Language/Model	Gujarati	Hindi	Marathi	Oriya	Tamil	Telugu	Avg
GMM-HMM	69.03	33.22	55.78	48.81	47.27	28.33	46.88
TDNN	40.41 (37.20)	22.44 (29.04)	39.06 (38.46)	33.35 (34.09)	30.62 (31.44)	19.27 (26.15)	30.73 (32.73)
TDNN (Monolingual)	18.23 (25.98)	31.39 (27.45)	18.61 (20.41)	35.36 (31.28)	34.78 (35.82)	28.71 (29.35)	27.85 (28.38)
TDNN-LSTM	15.15 (26.03)	40.45 (39.01)	22.04 (38.76)	38.95 (40.69)	33.75 (34.24)	30.81 (31.59)	30.19 (35.05)
TDNNF	20.05	41.92	23.97	40.17	33.58	32.12	31.96
B0	28.5	40.1	19	38	33	34.5	32.18
+ RNNLM	25.6	27.1	18.7	38.2	28.8	29.1	27.91
B1	24.9	26.6	18.6	37.0	29.1	29.5	27.61
+ RNNLM	23.2	24.6	17.9	33.4	25.7	25.7	25.08
+ 5-gram LM	19.3 **(34.57)**	23.5 **(21.49)**	16.7 **(46.41)**	33.0 **(32.13)**	23.3 **(28.6)**	21.9 **(28.03)**	22.95 **(31.87)**
+Clean	(42.44)	(27.7)	(49.29)	(36.11)	(36.48)	(37.88)	(38.32)
C0	26.5	28.1	18.8	40.8	30.7	32.1	29.5
+Clean	(83.56)	(41.91)	(79.65)	(67.66)	(58.29)	(62.0)	(65.51)
B3	51.8	37.0	55.8	67.9	94.1	88.6	65.86
B3 (transliterated)	31.2	33.0	19.7	37.8	43.9	40.3	34.31
C1	25.6	26.2	18.0	37.9	30.0	31.5	28.2
+ RNNLM	22.6	23.3	16.5	36.3	26.3	26.7	25.28
+ 5-gram LM	18.6 (37.5)	22.0 (22.64)	15.4 (33.87)	36.2 (38.74)	22.4 (30.84)	22.2 (33.61)	22.79 (33.87)

Channel distortion

- Odia - Present in train, dev and test set.
- Marathi - Absent in train, dev. Present in most of the examples in test set.
- Other languages - Absent in train dev. Present in a few examples in test set.





Effect of channel characteristics on different LID methods

Similar to 1-pixel attack (Su et al.)

Artificially augment the dev set with the noise from Odia data

B0 - Language independent ASR

C0 - Joint LID/ASR

L0 - LID on B0 Encoder

L1 - LID on Transliterated Encoder (B3)

dialpad

Workaround - retrain with clean data and clean test data

Language/Model	Gujarati	Hindi	Marathi	Oriya	Tamil	Telugu	Avg
GMM-HMM	69.03	33.22	55.78	48.81	47.27	28.33	46.88
TDNN	40.41 (37.20)	22.44 (29.04)	39.06 (38.46)	33.35 (34.09)	30.62 (31.44)	19.27 (26.15)	30.73 (32.73)
TDNN (Monolingual)	18.23 (25.98)	31.39 (27.45)	18.61 (20.41)	35.36 (31.28)	34.78 (35.82)	28.71 (29.35)	27.85 (28.38)
TDNN-LSTM	15.15 (26.03)	40.45 (39.01)	22.04 (38.76)	38.95 (40.69)	33.75 (34.24)	30.81 (31.59)	30.19 (35.05)
TDNNF	20.05	41.92	23.97	40.17	33.58	32.12	31.96
B0	28.5	40.1	19	38	33	34.5	32.18
+ RNNLM	25.6	27.1	18.7	38.2	28.8	29.1	27.91
B1	24.9	26.6	18.6	37.0	29.1	29.5	27.61
+ RNNLM	23.2	24.6	17.9	33.4	25.7	25.7	25.08
+ 5-gram LM	19.3 **(34.57)**	23.5 **(21.49)**	16.7 **(46.41)**	33.0 **(32.13)**	23.3 **(28.6)**	21.9 **(28.03)**	22.95 **(31.87)**
+Clean	(42.44)	(27.7)	(49.29)	(36.11)	(36.48)	(37.88)	(38.32)
C0	26.5	28.1	18.8	40.8	30.7	32.1	29.5
+Clean	(83.56)	(41.91)	(79.65)	(67.66)	(58.29)	(62.0)	(65.51)
B3	51.8	37.0	55.8	67.9	94.1	88.6	65.86
B3 (transliterated)	31.2	33.0	19.7	37.8	43.9	40.3	34.31
C1	25.6	26.2	18.0	37.9	30.0	31.5	28.2
+ RNNLM	22.6	23.3	16.5	36.3	26.3	26.7	25.28
+ 5-gram LM	18.6 (37.5)	22.0 (22.64)	15.4 (33.87)	36.2 (38.74)	22.4 (30.84)	22.2 (33.61)	22.79 (33.87)

Conclusion

- Compared implicit, explicit and joint LID+ASR models as well as hybrid models
- Fine-tuning a multilingual Encoder on language-specific decoders help
- Word-level n-gram LM helps to select correct form of word segmentation (solving agglutination problem)
- Joint LID/ASR models are sensitive to channel characteristics for the LID task and not using the phonetics of the languages for classification
- We call for a more careful analysis of the joint LID+ASR methods under noisy conditions
- Recipe and pre-trained models are open-sourced https://github.com/dialpad/mucs_2021_dialpad

dialpad

Demo