Interspeech 2022
18-22 September 2022, Incheon, Korea

Improving Language Identification of Accented Speech

Kunnar Kukk, Tanel Alumäe

Laboratory of Language Technology
Tallinn University of Technology, Estonia
{kunkuk, tanel.alumae}@taltech.ee

Abstract

Language identification from speech is a common preprocessing step in many spoken language processing systems. In recent years, this field has seen fast progress, mostly due to the use of self-supervised models pretrained on multilingual data and the use of large training corpora. This paper shows that for speech with a non-native or regional accent, the accuracy of spoken language identification systems drops dramatically, and that the accuracy of identifying the language is inversely correlated with the strength of the accent. We also show that using the output of a lexicon-free speech recognition system of the particular language helps to improve language identification performance on accented speech by a large margin, without sacrificing accuracy on native speech. We obtain relative error rate reductions ranging from to 35 to 63% over the state-of-the-art model across several non-native speech datasets.

Index Terms: language identification, non-native accent, bias

1. Introduction

Spoken language identification (LID) is the task of automatically identifying the language of an utterance. LID is often used as a preprocessing step in speech-based multilingual applications, such as spoken translation, human-machine communication systems and multilingual speech transcription systems. LID is also commonly used in automatic call routing where it is used to direct a call to a fluent native operator [1].

Since LID and ASR are increasingly used in critical services, it is important that such systems work flawlessly across a wide user community, with respect to variability corresponding to societally sensitive characteristics or traditionally marginalized communities, such as gender, ethnicity, disability, etc. Decreased robustness of speech based systems towards certain user groups may amplify biases already present in the society.

Speaking in a language other than one's native tongue is an ubiquitous reality in the globalized world. For example, in 2019, the number of international migrants was estimated to be 272 million, corresponding to 3.5% of the world's population [2]. The worldwide increase in the number of non-native speakers is caused both by rise in movement due to labour, study and leisure, but also by the large regional conflicts that cause a sudden increase in the number of refugees in certain parts of the world. Learning the local language is considered one of the most important aspects for migrants' inclusion in the society by both the receiving community and migrants themselves [3]. A large proportion of migrants are eager to learn the language of the receiving society at least to some degree, as it is important for helping migrants navigate a new environment, including access to health care, banking and other critical services. It also improves their access to education and employment [4]. However, almost 90% of the first generation and around 50% of the second generation migrants speak the local language with a weak or strong non-native accent [5]. Accent is not a phenomena that is specific to non-native speakers. Also native speakers can have a strong accent peculiar to a particular location or ethnicity, that is different from what is often regarded as a standard pronunciation. However, according to sociolinguistic approach, everyone has an accent, even the native speakers [6].

In recent years, the field of language identification from speech has seen a fast progress. This is mostly due to the use of self-supervised models trained on very large multilingual datasets (such as XLS-R [7]) and the emergence of large multilingual speech datasets with language identification labels, such as Mozilla CommonVoice [8] and VoxLingua107 [9]. The resulting models can achieve very high language identification accuracies on spoken data that contain mostly native speech. The recent Oriental Language Recognition 2021 Challenge [10] included a 17 language identification task with utterances obtained from real-life environments. The top performing teams [11, 12] achieved equal error rates below 1%. This might suggest that LID is a task that is close to be solved.

Several studies have shown that ASR systems produce more errors for non-native speech than for native speech [13, 14, 15]. This is not surprising, since ASR systems are usually trained on speech originating mostly from native speakers. In [16], it was discovered that non-native accent causes on the average three times more LID errors than native speech, when using phonotactic models.

The first goal of this paper is to quantify the accuracy of LID on no-native speech with state-of-the-art models that produce excellent results on native speech. We do this by measuring the performance of different LID models on datasets that contain non-native speech and compare the results with similar datasets containing native speech of the same language. We show that LID systems that provide excellent accuracy for native speech can degrade dramatically in the presence of non-native and regional accents. Then, we investigate using recognition hypotheses of one or many lexicon-free ASR systems as additional features when producing the LID decision. The idea of using ASR hypotheses for improving LID systems is not entirely new: both [17] and [18] experimented with combining acoustic and ASR-based features for improving LID and report around 50% relative reduction in error rate over the baseline acoustic model. We show that combining character n-gram based Naïve Bayes text classification models with a system that uses acoustic representations increases the robustness of LID systems to accented speech by a large margin, without sacrificing accuracy on native speech.

2. Experimental set-up

2.1. Datasets

The following section gives an overview of the 6 datasets used in this work. Table 1 summarises characteristics of datasets.

10.21437/Interspeech.2022-10455

Table 1: *Characteristics of datasets used for training and evaluation.*

Dataset	**Language**	**Non-native accent?**	**Sampling Rate (kHz)**	**Type**	**Utterances**	**Utterance Avg Length (sec)**
Estonian Foreign Accent Corpus	Estonian	Yes/No	44.1	Spontaneous/Dictated	32649	5.9
CSLU Foreign Accented English	English	Yes	8	Spontaneous	4925	17.9
CSLU 22 Languages (English)	English	No	8	Spontaneous/Dictated	2206	6.4
CMU Arctic	English	No	16	Dictated	14471	3.2
L2 Arctic	English	Yes	44.1	Dictated	25758	3.7
VoxLingua107 train	107	No	16	Spontaneous	2.54M	9.4
VoxLingua107 dev	33	No	16	Spontaneous	1608	10.0

2.1.1. *Estonian Foreign Accent Corpus*

Estonian Foreign Accent Corpus (EFAC, version 1) consists of speech data from 185 non-native (L2) and 20 native Estonian speakers. It contains 25-30 minutes of speech from each speaker [19, 20]. Speech is recorded in a studio, using is 16-bit 44.1 kHz stereo format. The dataset consists of 32649 utterances, totalling in 53.2h of speech (48.8h non-native and 3.4h native speech) with average utterance length of 5.9 seconds. Figure 1 shows distribution of utterance lengths of each dataset.

EFAC speech corpus contains examples of spontaneous and read speech (136 phonetically rich sentences and two short texts). The text corpus involves 130 neutral sentences including the main phonological oppositions of Estonian, eight questions, two passages, and prompts to elicit spontaneous speech (self-introduction, description of three pictures). The dataset also contains the subjects' self-assessment with regard to their Estonian proficiency level.

2.1.2. *CSLU Foreign Accented English*

CSLU Foreign Accented English (CSLU FAE) release 1.2 [21] consists of non-native speech in English by native speakers of 22 different languages. Speech is recorded via a telephone channel, using 16-bit 8kHz mono format. It includes spontaneous telephone speech, information about the speakers' linguistic backgrounds and perceptual judgments about the accents in the utterances. The speakers were asked to speak about themselves in English for 20 seconds, having an average utterance length of 17.9 seconds with a total 24 hours of 4925 telephone-quality utterances.

2.1.3. *CSLU 22 Languages English Subset*

The English subset of the CSLU 22 Languages Corpus [22] is a 3.9 hour dataset of native English. It contains speech recorded using 16-bit 8kHz mono format. Utterances contain both fixed vocabulary words as well as fluent continuous telephone speech. The number of utterances in the dataset is 2206, having an average utterance length of 6.4 seconds. As seen on Figure 1, a large proportion of the utterances are shorter than one second.

2.1.4. *CMU Arctic*

CMU Arctic dataset [23] consists of 12.9 hours of dictated speech from 18 native English speakers with various American accents as well as Canadian, Scottish and Indian accents. It contains 16-bit 16kHz mono-formatted data. Total number of utterances is 14471 with average utterance length of 3.2 seconds.

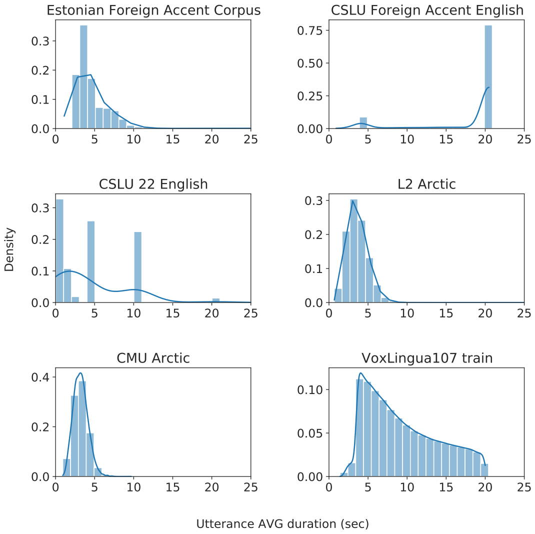

Figure 1: *Density of dataset utterance average duration.*

2.1.5. *L2 Arctic*

L2 Arctic dataset [24] consists of 26.4 hours of dictated English speech from 24 non-native speakers with 18 different language backgrounds, with an average of 67.7 minutes of speech per speaker. Speech is recorded using 16-bit 44.1 kHz mono format.The total number of utterances is 25758, with an average utterance length of 3.7 sec.

2.1.6. *VoxLingua107*

VoxLingua107 train set consists of 6628 hours of speech extracted from automatically scraped Youtube videos [9]. The language labels of the utterances are based on the detected language of the title and description of the video. Data-driven post-filtering was used to remove segments from the database that were likely not in the given language, increasing the proportion of correctly labeled segments in the dataset to 98%, based on crowd-sourced verification. VoxLingua107 has speech data across 107 different languages. Number of utterances in the dataset is about 2.54 million. The average utterance length is 9.4 seconds. The average amount of data per language is 62 hours.

VoxLingua107 development set consists of 4.5 hours of spontaneous speech from Youtube videos. It has speech data across 33 different languages. The language of the utterances in the development set has been verified by at least two native or proficient crowd-sourced speakers. Number of utterances in the dataset is 1608.

2.2. Methodology

In order to assess the impact of foreign accent on LID performance, we train several LID models on the VoxLingua107 training data and test their accuracy on the English and Estonian test sets and the VoxLingua107 development set. Our first goal is to find out how well current state-of-the-art LID models perform on non-native speech. Our second goal is to propose methods for improving the accuracy of LID on foreign-accented speech, without using any additional training data nor changing the priors of the model, and also without reducing the accuracy of LID on native speech.

When analyzing the accuracy of different LID models across the presented datasets, it is important to understand that there are many factors besides accent that impact the results, such as the length of the utterance, noise level and type of speech. The native and non-native subsets in EFAC are directly comparable, since they originate from the same dataset. Similarly, the data in CMU Arctic is very similar most aspects to that of L2 Arctic. However, the data in CSLU Foreign Accented English is quite different from the CSLU 22 Languages corpus. Although they both contain telephone speech, a large proportion of the utterances in the CSLU Foreign Accent English corpus are around 20 seconds in length, while the utterances in CSLU 22 English corpus are shorter, with many less than one second in length, which is expected to be very challenging for LID models.

2.3. Models

2.3.1. Resnet model

The Resnet-style model is derived from the x-vector paradigm [25, 26], with several enhancements. For frame-level feature extraction, we use the Resnet34 [27, 28] architecture where the basic convolutional blocks with residual connections are replaced with squeeze-and-excitation modules [29, 30]. The statistics pooling layer that maps frame-level features to segment level features is replaced in our model with a multi-head attention layer [31, 32]. The utterance-level features resulting from the attention-based statstics pooling layer are further processed by two fully connected layers that also apply batch normalization and the ReLU non-linearity. The model is trained using cross-entropy loss. The details of this model can be found in [12].

For LID, this model is not applied directly but it is used for extracting the embeddings for the training and test data. Embeddings are extracted from the output of the first fully connected layer that comes after pooling. The embeddings are centered on the training data and reduced to 108-dimensional vectors using Linear Discriminant Analysis (LDA). The final scoring is done using a Probablistic Linear Discriminant Analysis (PLDA) model.

2.3.2. XLS-R 300M

We also experimented with using the XLS-R-300M wav2vec2.0 model [7] as the backbone of our language embedding model. XLS-R-300M is trained on unlabeled multilingual data. The model is trained by jointly solving a contrastive task over masked latent speech representations and learning a quantization of the latents shared across languages. XLS-R is pretrained on around 500K hours of speech data from 128 languages.

We used XLS-R-300M as follows: the outputs from the wav2vec2 model were fed through an attentive pooling layer, a fully connected layer with ReLU and batch normalization, and the final output layer, corresponding to the languages of the training set. During training, the learning rate corresponding to the XLS-R model was set to 0.01 times lower than the base learning rate. As the final classification backend, similar LDA/PLDA based setup as for the Resnet model was applied. This model achieves state-of-the art results on the VoxLingua107 development set. It was also the main component of the system that was ranked 2nd in the unconstrained task of the OLR 2021 Challenge [12].

2.3.3. Multinominal Naïve Bayes model on ASR output

The multinomial Naïve Bayes (NB) model predicts the language of an utterance based on its ASR-based transcript. It uses word-internal character 4-grams as features, with n-grams at the edges of words padded with space. The model considers all n-grams that occur in training data and uses Laplace smoothing with the smoothing parameter set to 0.95.

For generating ASR transcripts, we used two models: English and Estonian. Both models are finetuned from the multilingual wav2vec2 models using connectionist temporal classification (CTC) loss. The English model[1] is finetuned from XLSR-53K [33] using the English CommonVoice data. The Estonian model[2] is finetuned from XLS-R-300M [7] using around 800 hours of diverse Estonian speech (mainly broadcast speech). Neither of those models use an external language model during decoding and both are using character-based vocabularies. This has two benefits: first, decoding using a GPU is very fast, making it feasible to decode the whole 6628 hours of VoxLingua107 data for generating training data. Second, the output of the lexicon-free ASR system is not constrained to in-vocabulary words, resulting in very expressive character ASR-transcripts for languages other than the ASR target language.

2.3.4. Convolutional Neural Network on ASR outputs

As an alternative to a NB-based text classification model, we experimented with a convolutional neural network (ConvNet). The proposed ConvNet consists of several parallel convolutional input branches that each process the ASR transcript generated using a particular ASR model. The outputs from convolutional input branches are pooled over the utterance using max-pooling, concatenated and further processed using two fully connected layers. The model is trained using cross-entropy loss. The convolutional branches first map characters to their learned 20-dimensional embeddings and then apply a series of convolutional layers. In our experiments, we used five 1D convolutional layers with kernel sizes $(3, 1, 3, 1)$, with the number of channels set to 512. Similarly to the acoustic-based LID models, the ConvNet model is not applied directly for inference but is used for extracting 512-dimensional embeddings (from the output of the first fully connected layer that comes after pooling). The embeddings are then processed using the LDA/PLDA model.

[1] `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`

[2] `https://huggingface.co/TalTechNLP/xls-r-300m-et`

Table 2: *Language identification accuracy of different models across the English and Estonian test sets.*

		English				Estonian		Various
		CMU Arctic	**L2 Arctic**	**CSLU FAE**	**CSLU 22 en**	**EFAC**	**EFAC**	**V107 dev**
ID	**Model**	**Native**	**Non-native**	**Non-native**	**Native**	**Native**	**Non-native**	**Native**
A	Resnet	77.4	60.5	67.1	57.7	93.3	43.9	91.9
B	XLS-R 300M	87.6	74.6	79.5	71.9	**99.6**	51.8	**95.3**
D	NB on en ASR char 4-grams	83.8	79.8	84.7	57.1	20.2	21.0	54.6
E	NB on et ASR char 4-grams	81.5	74.6	45.7	34.3	71.0	65.3	48.7
F	Fusion of D, E	91.9	88.2	81.7	58.7	68.2	62.8	58.5
G	LDA+PLDA on log probs of F	90.1	86.0	83.4	53.8	69.2	63.7	75.3
H	ConvNet on et+en ASR output	85.7	81.5	86.9	46.8	56.2	50.7	71.4
I	Fusion of B, G	**95.5**	**90.5**	**88.2**	**72.6**	99.5	**69.5**	**95.3**

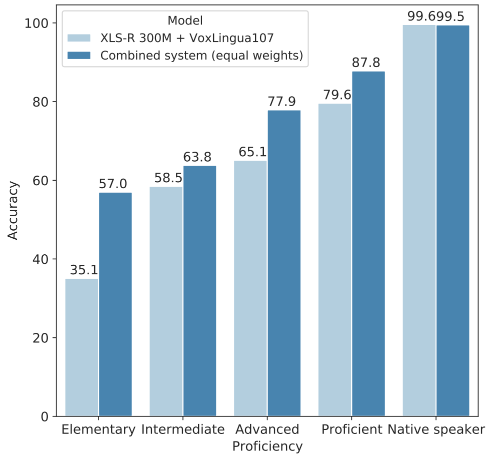

Figure 2: *Accuracy of identifying Estonian, depending on the speaker's self-estimated proficiency.*

3. Results and analysis

The results of different models and their combinations are listed in Table 2. All models are trained on the VoxLingua107 training set. The first two models (A and B) are based on only acoustic representations. It can be seen that using the finetuned XLS-R wav2vec2 model results in large improvements over the Resnet model on all datasets, regardless of the accent. The result on the VoxLingua107 development set is better than the previous state-of-the-art of 94.3% [7]. At the first sight it is surprising that the accuracy on the native English CMU Arctic dataset is much lower than the accuracy on the native Estonian subset of EFAC. Upon deeper inspection, it turns out that the accuracy varies a lot across the different speakers in CMU Arctic, ranging from 35% for a speaker with a distinctive Indian accent to 100% for a speaker without any marked pronunciation features (using model B). This indicates that the LID models using acoustic representations not only struggle with non-native speech, but also with native speech with a distinctive regional accent.

Models D-H are all based on ASR transcriptions. By comparing models D (Naïve Bayes model using character 4-grams from transcriptions generated using the English model) and E (same, but using Estonian model transcriptions) it can be seen that having the target language ASR transcripts available is more helpful for LID than other ASR transcripts: e.g., the accuracy of a model trained on the output of the Estonian ASR system (model D) achieves 71% accuracy on native Estonian, dropping to only 20% when using English ASR transcripts. Fusion of individual NB models using linear interpolation results in gains in LID performance for all datasets. Using the log posterior probabilities of the fused NB model as input to a LDA/PLDA based LID system results in further improvements for Estonian and VoxLingua107 dev set. Surprisingly, the ConvNet trained on ASR transcripts is not able to outperform the fusion of NB models.

Model I, the fusion of the best acoustic and ASR-based models, outperforms all models on most datasets by a large margin. The fusion uses uniform weights for the two models. For good performance on accented speech it is important not to optimize the fusion weights on native speech data (such as VoxLingua107): on native speech, acoustic representations result in much higher accuracy than the ASR-based features and the optimized model collapses into an acoustic-only model.

Figure 2 compares the accuracy of the acoustic model (B) and the fused model (I) on the EFAC data, using subjects' self-estimated proficiency to group the speakers. The chart confirms that there is a strong inverse correlation between the strength of the accent and the LID performance. Using ASR transcripts as additional features improves LID results across all proficiency levels for non-native speech. However, there is still a noticeable gap between LID accuracy of native and non-native speech, even for proficient non-native speakers.

4. Conclusion

In this paper, we have shown that LID systems that perform exceptionally well on native speech, have dramatically worse accuracy on identifying the language from non-native speech and native speech with a distinctive regional accent.

Experiments showed that this problem can be mitigated (but not fully solved) by using a LID model that fuses the predictions of the acoustic-based model with the outputs of a text classification model trained on the transcripts of one or many monolingual lexicon-free ASR systems. In our experiments, this helped to improve LID accuracy on non-native speech by a large margin, with relative error rate reductions ranging from to 35 to 63% over the state-of-the-art acoustic model, without decreasing accuracy of LID on native speech.

5. References

[1] "Language identification (LID). Phonexia Speech Platform," https://partner.phonexia.com/kb/sp/speech-platform/spe/technologies-available-spe/language-identification/, accessed: 2022-06-27.

[2] *World Migration Report 2020*. International Organization for Migration, 2019. [Online]. Available: https://publications.iom.int/books/world-migration-report-2020

[3] S. Castles, "Migration and community formation under conditions of globalization," *The International Migration Review*, vol. 36, no. 4, pp. 1143–1168, 2002.

[4] B. R. Chiswick, "Tongue tide: The economics of language offers important lessons for how Europe can best integrate migrants," *Finance & Development*, vol. 53, no. 003, 2016.

[5] I. Kogan, J. Dollmann, and M. Weißmann, "In the ear of the listener: The role of foreign accent in interethnic friendships and partnerships," *International Migration Review*, vol. 55, no. 3, pp. 746–784, 2021.

[6] M. J. Matsuda, "Voices of America: Accent, antidiscrimination law, and a jurisprudence for the last reconstruction," *The Yale Law Journal*, vol. 100, no. 5, pp. 1329–1407, 1991. [Online]. Available: http://www.jstor.org/stable/796694

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *ArXiv preprint*, vol. abs/2111.09296, 2021. [Online]. Available: https://arxiv.org/abs/2111.09296

[8] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, Marseille, France, 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[9] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[10] B. Wang, W. Hu, J. Li, Y. Zhi, Z. Li, Q. Hong, L. Li, D. Wang, L. Song, and C. Yang, "OLR 2021 challenge: Datasets, rules and baselines," *ArXiv preprint*, vol. abs/2107.11113, 2021. [Online]. Available: https://arxiv.org/abs/2107.11113

[11] A. Lyu and Z. Wang, "Ant multilingual identification system for OLR 2021," Tech. Rep., 2021.

[12] T. Alumäe and K. Kukk, "Pretraining approaches for spoken language recognition: TalTech submission to the OLR 2021 Challenge," in *Speaker Odyssey*, 2022.

[13] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *ArXiv preprint*, vol. abs/2103.15122, 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[14] Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan, "See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–9.

[15] A. Awasthi, A. Kansal, S. Sarawagi, and P. Jyothi, "Error-driven fixed-budget ASR personalization for accented speakers," in *ICASSP*, 2021, pp. 7033–7037.

[16] R. Wanneroy, E. Bilinski, C. Barras, M. Adda-Decker, and E. Geoffrois, "Acoustic-phonetic modeling of non-native speech for language identification," Multi-Lingual Interoperability in Speech Technology, Tech. Rep., 2000.

[17] S. Wang, L. Wan, Y. Yu, and I. L. Moreno, "Signal combination for language identification," *ArXiv preprint*, vol. abs/1910.09687, 2019. [Online]. Available: https://arxiv.org/abs/1910.09687

[18] C. Chandak, Z. Raeesy, A. Rastrow, Y. Liu, X. Huang, S. Wang, D. K. Joo, and R. Maas, "Streaming language identification using combination of acoustic representations and ASR hypotheses," *ArXiv preprint*, vol. abs/2006.00703, 2020. [Online]. Available: https://arxiv.org/abs/2006.00703

[19] L. Meister and E. Meister, "Aktsendikorpus ja võõrkeele aktsendi uurimine," *Keel ja Kirjandus*, vol. 55, no. 8-9, pp. 696–714, 2012.

[20] "Estonian Foreign Accent Corpus," https://doi.org/10.15155/9-00-0000-0000-0000-0002BL, accessed: 2022-03-21.

[21] T. Lander, *CSLU: Foreign Accent English Release 1.2 (LDC2007S08)*. Linguistic Data Consortium, 2007.

[22] ——, *CSLU: 22 Languages Corpus (LDC2005S26)*. Linguistic Data Consortium, 2005.

[23] J. Kominek and A. W. Black, "CMU Arctic databases for speech synthesis," Language Technologies Institute School of Computer Science Carnegie Mellon University, Tech. Rep., 2003.

[24] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," in *Interspeech*, 2018, pp. 2783–2787. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1110

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461375

[26] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[27] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[30] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function," in *Interspeech*, 2019, pp. 2883–2887. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1704

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[32] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Interspeech*. ISCA, 2018, pp. 3573–3577. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1158

[33] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech*, 2021. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329