INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Embedding Articulatory Constraints for Low-resource Speech Recognition Based on Large Pre-trained Model

Jaeyoung Lee, Masato Mimura, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan
{jaeyoung, mimura, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Knowledge about phonemes and their articulatory attributes can help improve automatic speech recognition (ASR) of low-resource languages. In this study, we propose a simple and effective approach to embed prior knowledge about phonemes into end-to-end ASR based on a large pre-trained model. An articulatory attribute prediction layer is constructed by embedding articulatory constraints in layer initialization, which allows for predicting articulatory attributes without the need for explicit training. The final ASR transcript is inferred by combining the output of this layer with encoded speech features. We apply our method to finetune a pre-trained XLS-R model using Ainu and Mboshi corpora, and achieve a 12% relative improvement when target data of only 1 hour is available. This demonstrates that the approach of incorporating phonetic prior knowledge is useful when combined with a large pre-trained model.

Index Terms: Low-resource speech recognition, articulatory attributes, wav2vec2.0

1. Introduction

Over the last decade, deep neural network (DNN) based approaches have significantly improved the performance of automatic speech recognition (ASR) and have made the end-to-end approach possible [1] [2] [3], where output prediction is directly inferred from acoustic features through a single neural network, without the need of manual pronunciation modeling. However, this has been possible with large language corpora, and only a handful of languages have sufficient language resources to achieve high performance for ASR, among approximately 7,000 languages [4] in the world. In particular, the number of nodes in the output layer in end-to-end models may be too large for low-resource languages, and this can be partially mitigated by encoding the relationships between output tokens.

Articulatory attributes are a set of distinct features that describe how speech sounds are produced by the articulators in the mouth, such as the lips, tongue, and vocal cords. It is shown that articulatory attributes can be recognized across different languages [5]. The approach incorporating articulatory attributes to ASR systems has been investigated for traditional GMM-HMM based models, and shown to contribute to making models more robust to speaker or channel variability [6]. In Automatic Speech Attribute Transcription (ASAT) [7] [8], a bank of speech attribute detectors were placed in the lowest level of ASR pipeline hierarchy, where detected attributes were combined to predict phones, syllables and words. Articulatory modeling has also been applied to DNN-HMM and end-to-end models in multilingual settings, improving robustness to spontaneous and non-native speech [9], and benefiting performance in low-resource scenarios [10] [11] [12].

In recent years, a prominent trend in low-resource speech recognition research has been to use self- or semi-supervised pre-training on high-resource speech corpora to learn a universal speech representation, which can then be finetuned for downstream tasks [3]. Large pre-trained multilingual models such as wav2vec2.0 and XLS-R [13] are shown to learn general representation that is applicable even to unseen languages, and have greatly benefited low-resource ASR performance, as shown in [14]. It is known that these models learn high-level representations corresponding to phonemes without any explicit supervision [3], and it is possible that part of the learned representation corresponds to articulatory attributes as well. Therefore, the approach of incorporating articulatory information is expected to be effective when used in combination with a large-scale pre-trained model for developing an ASR system with very low-resource settings.

In this study, we propose a simple method to incorporate articulatory information by embedding it into layer initialization in end-to-end ASR. First, we construct a fixed-length encoding vector for each phoneme, using knowledge about articulatory attributes. Then, these attribute vectors are stacked to form an *articulatory attribute projection matrix*, which projects articulatory attribute prediction into output phoneme prediction. This articulatory attribute prediction layer is combined with another conventional projection layer to generate final outputs. We finetune a pre-trained XLS-R model with this output layer placed on top. We also explore multilingual training that exploits high-resource language to enhance the representation ability of the model for articulatory attributes. The proposed method is applied to two low-resource languages of Ainu and Mboshi, with the target training data of around 34 and 4 hours, respectively.

2. Related Work

2.1. Articulatory Modeling

Müller et al. [11] attempted to improve low-resource ASR by using articulatory attributes along with language feature vectors and acoustic features. They introduced seven articulatory attribute classifiers and one special phoneme type classifier for each of the 8 categories they defined, such as *phoneme type, manner and place of articulation*, and *vowel frontness*. Each category has a predefined number of classes, including a special class representing *not applicable*; for example, *frontness* category has 4 classes: *front, central, back, not applicable*. This modeling is rather restrictive because each phoneme can only have one attribute class for each category, making it difficult to apply to multilingual settings with languages requiring different articulatory modeling. Moreover, it was not implemented in an end-to-end ASR model.

Li et al. [12] introduced articulatory modeling to end-to-

10.21437/Interspeech.2023-1437

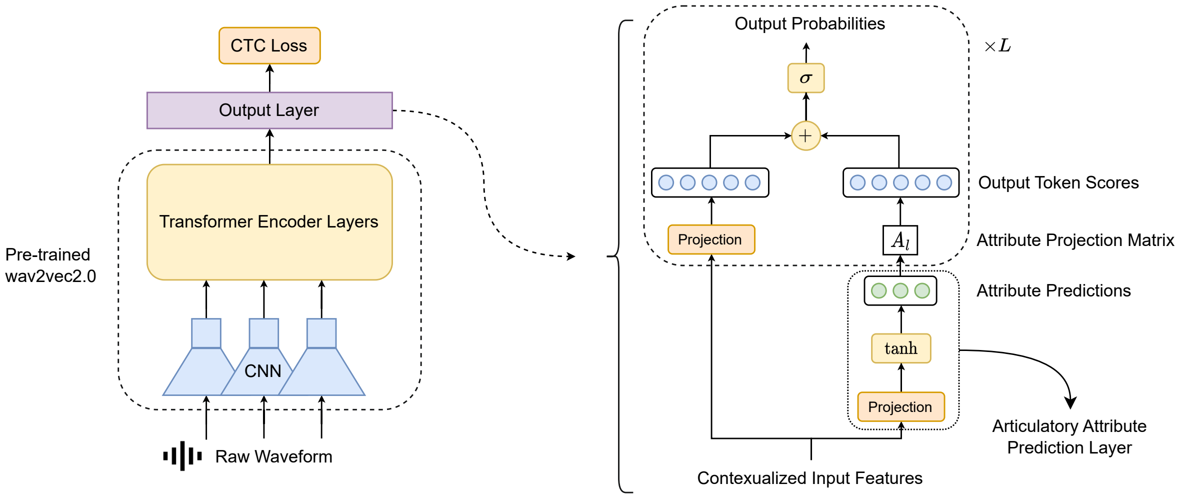

Figure 1: *An overview of the proposed model architecture. Each language has a respective output layer, while the articulatory attribute prediction layer is shared across languages. L stands for the number of languages. A_l stands for the articulatory attribute projection matrix for language l and its elements are set as trainable parameters and is implemented simply as a linear projection layer.*

end speech recognition. They mapped each character token to a sequence of tokens representing articulatory attributes. For example, the character representing */g/* is mapped to the sequence of two tokens ⟨*voiced*⟩ ⟨*velar*⟩. They trained a Transformer-based model to predict articulatory attribute token sequences from speech features in an end-to-end fashion. They observed that while this method underperforms under monolingual settings, it can significantly outperform usual end-to-end models in multilingual settings, where the model can effectively learn articulatory representations shared across target languages. This suggests the importance of multilingual training when using articulatory features for speech recognition.

In this study, we incorporate articulatory attribute predictors and combine their outputs with acoustic features to form the final token prediction. Unlike [11], we treat each articulatory attribute independently, allowing phonemes to be modeled to have multiple attributes at the same time, e.g. affricates can be modeled by both being plosive and fricative. Moreover, we do not train such predictors in a supervised manner; instead, they are *induced* to make articulatory attribute predictions, by layer initialization. Then, the entire network is finetuned in an end-to-end manner. The modeling is more flexible to model attributes with a continuous scale, e.g. vowel height, and it can work with imperfect knowledge about the phonemes of the target language and inaccurate articulations in input speech.

2.2. Wav2vec2.0

Wav2vec 2.0 is a self-supervised learning framework that learns latent representation from raw speech data. In this framework, the speech input is first encoded by a multi-layer convolutional neural network (CNN). This latent representation is masked and input to a Transformer encoder network, producing contextualized representations. The model is trained through predicting true latent representation from other contextualized representations, in a similar fashion to masked language models such as BERT [15]. Vector quantized codes are used as a similarity measure. Wav2vec 2.0 models can be trained on large multilingual corpora [13] and can then be used as a pre-trained model for finetuning on a small amount of labeled data. It has been shown to outperform previous approaches especially in terms of low-resource ASR [14].

3. Proposed Method

3.1. Articulatory Modeling for Phonemes

First, we identify all relevant articulatory attributes for the target language, which are categorized into three groups: vowel attributes, place and manner of articulation for consonants. Additionally, there is a special category of attributes that applies to all output tokens. Table 1 lists all of the attributes covered in this study, for Ainu and Mboshi. Since different languages have different contrasting features, the set of articulatory attributes should be modified to encompass all contrasting features present in other target languages.

Vowel attributes (e.g. *vowel height, backness, etc*) tend to be best represented in a continuous scale, while consonant attributes (i.e. *place or manner of articulation*) tend to be best

Table 1: *Articulatory attributes.*

Category	Attributes		
	1	0	-1
Special	sound	-	symbol
	consonant	semi-vowel	vowel
	voiced	-	voiceless
Vowel	back	central	front
	open	mid	closed
	high-toned	-	low-toned
Consonant (place)	bilabial	-	else
	labiodental	-	else
	alveolar	-	else
	palatal	-	else
	velar	-	else
	glottal	-	else
Consonant (manner)	nasal	-	else
	plosive	-	else
	fricative	-	else
	flap	-	else
	approximant	-	else

Table 2: *Examples of assignment of articulatory attribute encoding. Encodings that deserve special attention are rendered bold.*

		sound / symbol	consonant / vowel	back / front	open / closed	voiced	plosive	fricative	bilabial	alveolar	palatal
	$\langle wb \rangle$	-1	0	0	0	0	0	0	0	0	0
Ainu	a	1	-1	0	1	1	0	0	0	0	0
	u	1	-1	1	-1	1	0	0	0	0	0
	c	1	1	0	0	**0**	**1**	**1**	-1	1	**0**
Japanese	f	1	1	0	0	-1	-1	1	1	-1	-1
	by	1	1	0	0	1	1	-1	1	-1	**1**

represented in a one-hot encoding scheme. In our method, we represent every attribute in a continuous scale ranging from -1 to 1, e.g. *front* (-1) to *back* (1). In the case of consonant attributes, a value of -1 denotes the attribute's absence, analogous to 0 in one-hot encoding. The interpretation of a value of 0 depends on the attribute and the token. For vowel attributes of vowel tokens, a value of 0 denotes the middle point on the -1 to 1 scale. For other instances, a value of 0 indicates that the attribute is not applicable, such as consonant attributes for vowel tokens.

We flexibly utilize this encoding scheme to embed phonetic and phonological knowledge, as it will only be used in layer initialization and not for training labels, and deep learning will finetune any incomplete, ambiguous or incorrect details. For example, the phoneme /c/ in Ainu is an affricate and is encoded by setting both *plosive* and *fricative* attributes to 1. The *palatal* attribute is set to 0, as it is sometimes realized as palato-alveolar [ʧ], though usually realized as alveolar affricate [ts]. Moreover, the *voiced* attribute is set to 0, encoding the fact that there is no phonemic contrast between voiced and voiceless consonants in Ainu. Table 2 shows examples of assignment of articulatory attributes for Ainu and Japanese phoneme tokens.

By representing each token as a fixed-length encoding vector, we can construct an articulatory attribute projection matrix $A_l \in \mathbb{R}^{V \times N}$ for a target language l, where V represents the size of vocabulary and N represents the number of articulatory attributes. Each row in A_l is normalized to have a unit variance and zero mean. This matrix can be regarded of as a mapping from [-1,1]-normalized attribute predictions to token predictions.

For languages with a restricted syllabic structure such as Ainu, it is straightforward to extend this encoding to syllabic tokens. Syllables in Ainu have (C)V(C) structure, having 3 phonemes at most. Thus, each syllable token in Ainu can be represented by a $3N$-dimensional encoding vector, and an attribute projection matrix can be constructed accordingly.

3.2. Articulatory Attribute Prediction Layer

As shown in the right-most part of Figure 1, the articulatory attribute prediction layer is placed to project input features to the attribute space using the $\tanh$ activation function, as we represent each attribute in a [-1,1] scale. This layer is followed by another projection layer from attribute space to output tokens, which is initialized by the attribute projection matrix A_l defined earlier. The outputs are then fed into softmax, yielding the token predictions. The attribute prediction layer is trained as a predictor for articulatory attributes without explicit supervision. However, we observed that using only the outputs from this layer as final token predictions leads to suboptimal performance. To address this, we combined it with conventional projection from input speech features to make the final predictions, as illustrated in Figure 1.

4. Experimental Setup

4.1. Datasets

4.1.1. Speech Corpus of Ainu Folklore

Ainu is a language spoken by the Ainu, a minority ethnic group in the northern part of Japan, and is classified critically endangered by UNESCO. The speech corpus of Ainu folklore [16] is a collection of speech recordings of Ainu stories, myths, and legends that have been collected to preserve the Ainu language and culture. It consists of utterances from 8 Ainu speakers speaking the Saru dialect, amounting to 38.9 hours. We split it into *train* set of 33.7 hours and *dev* set of 5.2 hours. Subsets of the *train* set with varying amounts of data are used for training, to investigate the effect of the amount of training data for our method. Furthermore, additional utterances of 14 hours spoken by another Ainu speaker with a distinct dialect, Shizunai dialect, is employed as the *test* set. The characters used in the transcription correspond 1-to-1 to phonemes.

In addition, we conduct experiments in bilingual settings where we adopt Japanese as an auxiliary language because Japanese and Ainu share most of the phonemes, and most Ainu speakers also speak Japanese. We use utterances of about 300 hours from the Corpus of Spontaneous Japanese (CSJ) [17] as an additional training corpus.

Both Ainu and Japanese have a restricted syllabic structure of (C)V(C), and thus it is straightforward to employ articulatory modeling for syllables in both languages. We use syllable as the target unit and employ syllabic articulatory modeling, as described in section 3.1.

4.1.2. Mboshi Parallel Corpus

Mboshi (Bantu C25, Congo-Brazzaville) is a Bantu language spoken by the Mboshi people in the Republic of Congo. The Mboshi parallel corpus [18] contains speech utterances of around 4.5 hours from 3 speakers. The data is split into *train* set of 3.9 hours and *dev* set of 26.4 minutes. We adopt the *train*/*dev* split as defined in the original paper [18].

The characters used in the transcription of this corpus do not correspond to phonemes in a 1-to-1 fashion. For example, the character *h* only appears in combination of either /gh/ or /bh/, which correspond to the voiced bilabial fricative and voiced velar fricative, respectively. To handle such cases, we leverage the flexibility of our articulatory modeling and assign a value of 1 to both the *bilabial* and *velar* attributes for *h*.

4.2. Model Training and Evaluation Measure

We used a publicly available pretrained multilingual model for all our experiments, namely XLS-R (0.3B) [13]. It is a 317M

Table 3: *CER(%) with varying amounts of Ainu training sets and Mboshi. The final row shows the relative improvement of the best performing proposed model from the baseline model.*

Target language (*unit*)	Ainu (*syll*)					Mboshi (*char*)
Training data	10m	1h	4h	10h	all (33.7h)	all (3.9h)
baseline	35.1	19.6	16.5	15.5	14.1	6.10
proposed (*attribute prediction only*)	26.0	18.0	15.8	15.2	14.1	7.28
proposed (*hybrid*)	24.8	17.6	**15.5**	14.8	**13.4**	**5.76**
proposed (*hybrid + bilingual training*)	**23.8**	**17.2**	15.6	**14.5**	14.1	-
relative improvement (%)	32.1	12.2	6.1	6.1	4.8	5.6

parameter model with 24 Transformer encoder layers, with embedding size of 1024 and 16 attention heads. It is trained on various speech corpora such as VoxPopuli, MLS, and CommonVoice, totaling 436K hours comprised of 128 different languages. We conduct finetuning experiments with CTC loss, where we freeze the convolution based feature extractor and finetune all of the transformer encoder layers, along with the output layers.

Models are trained using the Adam optimizer [19]. Learning rate is linearly warmed up for the first 10% training steps and peaks at 5e-3 for Ainu and 6e-3 for Mboshi, holds for 40% of the training steps, then gets exponentially decayed. The batch size is 45s and we mix target and auxiliary languages in 1:1 ratio in bilingual settings. We employ different numbers of training steps depending on the amount of target training data: 16K, 20K, 30K, 40K, 50K steps for training data of 10m, 1h, 4h, 10h, 33.7h, respectively. For bilingual training, we use 60% more training steps.

Three different configurations of the proposed method are trained and evaluated, and compared to the baseline configuration where the output layer simply consists of one linear projection layer [1]. In *attribute prediction only* configuration, only output from the articulatory attribute prediction layer is used for the final prediction. It is combined with input speech features in *hybrid* configuration, as illustrated in the right-hand side of Figure 1. In *hybrid + bilingual training* configuration for Ainu, the model is trained in a bilingual setting with additional training data in Japanese.

Character error rate (CER) is employed as the evaluation metric, given the difficulty in computing word error rate (WER) for very low-resource languages where determining word units is not obvious. For Ainu, the model with the lowest CER against *dev* set is selected and evaluated against the *test* set. For Mboshi, we simply evaluated the model at the end of the training against the *dev* set.

5. Experimental Results

The comparison of Character Error Rate (CER) of the baseline and proposed models in various training settings is presented in Table 3. We observe a significant improvement in performance in all *hybrid* models, especially when the available training data is smaller. The *attribute prediction only* models perform consistently worse than *hybrid* models, sometimes even worse than the baseline model. It is likely because the articulatory constraints can be too strict for the model to learn accurate representation of output tokens. The use of bilingual data improves performance by a small margin, although not significantly, when the training data is less than 10 hours. However, bilingual training with the entire 33.7h Ainu training data results in a decline in performance. This may be because we are modeling articulatory attributes of different languages in the same way, even though they are physically realized in a slightly different way.

Table 4: *CER(%) results with different inputs for articulatory attribute prediction layer*

Transformer Encoder #	Ainu (4h)
24	**15.5**
16	16.0
8	16.2
CNN features	16.5

Our method benefited performance for Mboshi as well as Ainu, where we observe 5.6% relative improvement. The absolute CERs for Mboshi are much lower than that of Ainu, and it is likely due to the fact that the speakers and dialects used in training and evaluation are the same, whereas the Ainu *test* set consists of a different speaker with a distinct dialect. The results demonstrate that out method can be applied and improve performance across different languages and different target units.

5.1. Effect of Inputs to the Attribute Prediction Layer

We conducted an additional series of experiments where we change the source of input to the articulatory attribute prediction layer. We tested latent representations from the final (24th), 16th and 8th Transformer encoder layer, as well as directly from the CNN feature extractor. Note that the final output of the Transformer encoder is still combined to produce the final output; with only input to the attribute prediction layer changed. The results are presented in Table 4. Notably, the results show a consistent trend in performance improvement with higher representation inputs. This finding suggests that articulatory attributes are high-level features that benefit from contextualized high-level representations.

6. Conclusions

We have presented a method for improving ASR in low-resource settings by incorporating linguistic prior knowledge about phonemes. It is simple yet flexible for modeling continuous and ambiguous articulatory constraints. It does not require explicit supervision on articulatory attribute labels and can model phonemes from multilingual inventories. We conducted finetuning experiments with XLS-R models on two very low-resource languages, Ainu and Mboshi. The results show that our method can improve CER by relative 12% when the amount of training data is limited to 1 hour, and still produce a meaningful improvement when using 33.7 hours of training data. This study demonstrates the importance of incorporating linguistic prior knowledge about target language, even when state-of-the-art pre-trained models are employed.

[1] The difference in model size between baseline and proposed configurations is marginal, as the pretrained XLS-R model already has 317M parameters.

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[4] M. P. Lewis, Ed., *Ethnologue: Languages of the World*, sixteenth ed. Dallas, TX, USA: SIL International, 2009.

[5] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 1, 2003.

[6] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[7] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. R. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in *Proc. Interspeech 2007*, 2007, pp. 1825–1828.

[8] C.-H.Lee and M.Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification and recognition," *Proc. IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[9] V. Mitra, W. Wang, C. Bartels, H. Franco, and D. Vergyri, "Articulatory Information and Multiview Features for Large Vocabulary Continuous Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5634–5638.

[10] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," vol. 36, no. C, pp. 173–195, 2016.

[11] M. Müller, S. Stüker, and A. Waibel, "Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features," in *Proceedings of the 13th International Conference on Spoken Language Translation*. International Workshop on Spoken Language Translation, 2016.

[12] S. Li, C. Ding, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "End-to-End Articulatory Attribute Modeling for Low-Resource Multilingual Speech Recognition," in *Interspeech 2019*. ISCA, 2019, pp. 2145–2149.

[13] 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," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[14] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages," 2021.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[16] K. Matsuura, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Speech Corpus of Ainu Folklore and End-to-end Speech Recognition for Ainu Language," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 2622–2628.

[17] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous Speech Corpus of Japanese," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA), 2000.

[18] A. Rialland, M. Adda-Decker, G.-N. Kouarata, G. Adda, L. Besacier, L. Lamel, E. Gauthier, P. Godard, and J. Cooper-Leavitt, "Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2000.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.