electronics

MDPI

Article

Latent Regression Bayesian Network for Speech Representation

Liang Xu [1,2], Yue Zhao [1,2,*], Xiaona Xu [1,2], Yigang Liu [1,2] and Qiang Ji [3]

[1] School of Information Engineering, Minzu University of China, Beijing 100081, China
[2] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China
[3] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA
* Correspondence: zhaoyueso@muc.edu.cn

Abstract: In this paper, we present a novel approach for speech representation using latent regression Bayesian networks (LRBN) to address the issue of poor performance in low-resource language speech systems. LRBN, a lightweight unsupervised learning model, learns data distribution and high-level features, unlike computationally expensive large models, such as Wav2vec 2.0. To evaluate the effectiveness of LRBN in learning speech representations, we conducted experiments on five different low-resource languages and applied them to two downstream tasks: phoneme classification and speech recognition. Our experimental results demonstrate that LRBN outperforms prevailing speech representation methods in both tasks, highlighting its potential in the realm of speech representation learning for low-resource languages.

Keywords: speech representation; latent regression Bayesian network; low-resource language

Citation: Xu, L.; Zhao, Y.; Xu, X.; Liu, Y.; Ji, Q. Latent Regression Bayesian Network for Speech Representation. *Electronics* **2023**, *12*, 3342. https://doi.org/10.3390/electronics12153342

Academic Editors: Jungpil Shin, Hoang D. Le and Md. Al Mehedi Hasan

Received: 28 June 2023
Revised: 3 August 2023
Accepted: 3 August 2023
Published: 4 August 2023

1. Introduction

Speech representation learning involves the utilization of machine learning techniques to extract abstract and high-level representations from speech signals. The objective is to transform the speech signal into a feature representation that encompasses various types of information, such as phonemes, emotions, and speaker characteristics. The extraction of effective speech representations from speech signals is crucial for reducing the learning complexity of downstream tasks and minimizing the amount of annotated data required. This holds importance in the realm of speech processing, particularly in the context of low-resource speech processing.

With over 7000 languages existing worldwide, many of them suffer from inadequate speech data, making it challenging to develop systems for low-resource speech tasks. Constructing such systems necessitates a resource-intensive data collection and labeling processes. To tackle this issue, researchers employ high-quality speech representation models, trained on extensive speech datasets, to serve as feature extractors by extracting audio features. Speech representation learning can be achieved through both supervised and unsupervised approaches. Supervised speech representation learning typically involves the utilization of large volumes of annotated data from high-resource languages for pre-training. The trained models are then employed as feature extractors, for instance, through the use of bottleneck features (BNF) [1–3]. However, the efficacy of pre-training is affected by the linguistic variability between the source and target tasks. On the other hand, unsupervised speech representation learning does not require the labeling of input data. This method relies on training models with abundant unannotated speech data and learning by reconstructing the input frames [4–7]. Additionally, self-supervised learning, which falls under the umbrella of unsupervised learning methods, leverages information constructed from unsupervised data itself as labels for learning representations that prove useful for downstream tasks [8]. One of the most advanced self-supervised speech representation methods is Wav2vec 2.0 [9], and large self-supervised models such as Wav2vec 2.0

have achieved human equivalence on numerous datasets. However, despite the promising results demonstrated by the Wav2vec 2.0 framework, its high computational cost poses a challenge when it comes to the practical application of these speech representations on low-resource languages [9]. Moreover, one study [10] has indicated that Wav2vec 2.0 is not universally applicable to all low-resource languages.

In this paper, our proposal centers around the utilization of a latent regression Bayesian network (LRBN) for speech representation learning in low-resource languages. The aim is to enhance the performance of low-resource speech systems. Unlike large models such as Wav2vec 2.0, LRBN is a lightweight model that learns data distribution and latent features in an unsupervised manner. By reconstructing its input through the latent layer, LRBN extracts the latent representation from the latent activation values. The goal is to ensure that the latent representation retains the essential features of the original data while being easier to analyze and process. This lightweight nature of LRBN is advantageous when dealing with small sample data, making it well-suited for speech representation learning in low-resource languages. Classical probabilistic deep generative models, such as restricted Boltzmann machines (RBM) and deep belief networks (DBN), often overlook correlations between latent variables, leading to a reduction in the representational power of the models [11]. In contrast, LRBN preserves the dependencies between latent variables. Studies have demonstrated that LRBN outperforms existing models in terms of data reconstruction and achieves comparable data representation performance [12]. To evaluate the effectiveness of LRBN in speech representation learning, we conducted experiments on two downstream tasks: phoneme classification and speech recognition. The results illustrate that LRBN surpasses previous approaches in these tasks, highlighting its efficacy in speech representation learning for low-resource languages.

Our research makes contributions in the following ways: (1) We extend the application of LRBN to address the issue of poor performance in speech recognition tasks for low-resource languages. LRBN, which has previously demonstrated success in the image domain for data representation, is introduced to speech representation learning in our work. (2) Through comprehensive experiments involving five different low-resource languages, including Tibetan, Cantonese, Korean, Uyghur, and Mongolian, we have demonstrated that our LRBN method achieves comparable or even superior performance against other speech representation techniques. (3) We investigated the impact of varying the number of latent layer nodes of LRBN on speech representation. Our findings suggest that optimizing the number of latent layer nodes can improve the effectiveness of our approach, highlighting its potential for further optimization.

2. Related Work

In the realm of speech representation, two distinct categories can be identified: traditional methods and deep learning methods. Traditional speech representation involves the use of manually crafted speech features, which offer computational simplicity, consistency, and high interpretability when compared to deep features. Traditional features such as Mel Frequency Cepstral Coefficients (MFCCs) and Filter Bank (FBank) are widely used features in speech processing. The process of extracting MFCCs involves several steps, including computing the Fourier transform, mapping the power spectrum to the Mel non-linear spectrum, and applying the discrete cosine transform (DCT) for decorrelation [13,14]. The latter step is essential for machine learning algorithms. Subsequently, FBank features were introduced. Unlike MFCCs, FBank features do not employ the DCT for decorrelation and offer superior information content with reduced computational complexity and higher feature correlation. Traditional features primarily focus on low-level attributes of speech, such as frequency, energy, and harmonics. In contrast, deep learning speech representation captures high-level attributes and the underlying relationships in speech signals, thereby enhancing the performance of low-resource speech systems for various languages [15]. Furthermore, studies have demonstrated that deep learning-based speech representations

exhibit superior generalization capabilities and do not necessitate features specifically designed for particular downstream tasks [14].

Deep learning-based speech representations can be acquired through supervised or unsupervised learning. Supervised learning allows for effective speech representations to be learned from large annotated datasets, such as Bottleneck Features (BNF), which is particularly useful for low-resource languages. In practice, the work of [16,17] has demonstrated that BNF can enhance the performance of low-resource speech tasks. However, if the pre-trained source language substantially differs from the target language of the task, the model may fail to capture the necessary acoustic and linguistic information required for the low-resource target task. On the other hand, unsupervised learning can be employed by training models with ample unannotated speech data, offering the potential to learn meaningful representations from speech [6,7]. Within the speech domain, a subcategory of unsupervised learning called self-supervised learning has garnered attention. Contrastive Predictive Coding (CPC), a self-supervised training criterion [8], facilitates representation learning by predicting adjacent frames based on the current frame. Wav2vec, introduced by Schneider et al. [18], directly employs the CPC loss for speech representation learning. Baevski et al. [19] proposed an improved method called Vq-wav2vec, which employs vector quantization to quantize and learn discrete speech representations from extracted features. Additionally, Baevski et al. [9] put forward Wav2vec 2.0, a combination of the Vq-wav2vec method and a masked language model (MLM) for training discrete speech units with contextual representations [20]. In addition to these methods, several recent theoretical and empirical works have focused on speech representation tasks. Refs. [21,22] delved into the representation and generalization capabilities of deep neural networks (DNNs) for speech processing. Notably, ref. [23] even explored the application of quantum neural networks to extract more representative speech features for low-resource speech recognition.

Presently, Wav2vec 2.0 stands as one of the most widely adopted speech feature representation methods in the domain of low-resource speech recognition. In ref. [24], Wav2vec 2.0 was utilized as a feature extractor for low-resource speech recognition. This study demonstrates the ability of Wav2vec 2.0 to learn high-quality feature representations from unlabeled speech data through self-supervised learning, which are then employed for speech recognition tasks. The results indicate that Wav2vec 2.0 exhibits remarkable transferability. Initially, the focus of Wav2vec 2.0 research was primarily on English. However, researchers in ref. [25] subsequently pre-trained a new version of Wav2vec 2.0 called XLSR using 56k hours of speech audio from 53 different languages. This model learns cross-language speech representations by pre-training on raw multilingual speech waveforms and demonstrates that cross-language pre-training can enhance speech recognition performance. While the Wav2vec 2.0 framework displays promising outcomes, it is computationally demanding [9]. Moreover, it has been noted in ref. [10] that Wav2vec 2.0 is not universally applicable to all low-resource languages. Thus, in this paper, LRBN is employed for low-resource speech representation learning to explore a more suitable approach for low-resource speech representation.

3. Methods

3.1. Latent Regression Bayesian Network

The LRBN model, utilizing the Bayesian network framework, enables the modeling of intricate interdependencies among variables. This model explicitly aims to capture the dependencies between latent variables for effective data representation. It comprises a visible layer X with n_d dimensions and a latent layer H with n_h dimensions, as depicted in Figure 1. Each latent variable is connected to visible variables through directed edges.

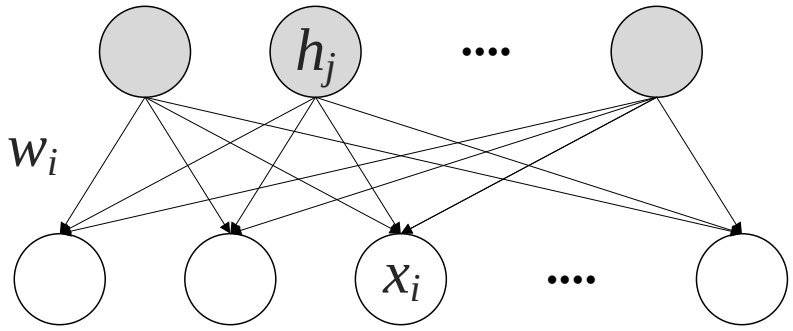

Figure 1. LRBN structure.

LRBN fulfills the chain rule, whereby it represents the visible variables as $\boldsymbol{x} = \left(x_1, \ldots, x_{n_d}\right)$ and the hidden variables as $\boldsymbol{h} = \left(h_1, \ldots, h_{n_h}\right)$. The joint probability distribution of all visible and hidden variables can be formulated as the product of the prior probability distribution of any hidden variable and the conditional probability distribution of any visible variable given the values of the hidden variables. The joint probability of x and h is computed in Equation (1):

$$P(\boldsymbol{x}, \boldsymbol{h}) = \prod_{j=1}^{n_h} P\left(h_j\right) \prod_{i=1}^{n_d} P(x_i \mid \boldsymbol{h}) \tag{1}$$

where n_h and n_d refer to the number of hidden and visible nodes, respectively. The joint probability of the visible and hidden variables is represented by $P(\boldsymbol{x}, \boldsymbol{h})$, whereas the prior probability of the hidden variable h_j is denoted as $P\left(h_j\right)$. Furthermore, $P(x_i \mid \boldsymbol{h})$ represents the conditional probability of the visible variable x_i given all hidden variables $\boldsymbol{h}$. Both $P\left(h_j\right)$ and $P(x_i \mid \boldsymbol{h})$ follow Bernoulli distribution and can be expressed as Equations (2) and (3), respectively:

$$P\left(h_j\right) = \sigma\left(d_j\right)^{h_j} \left(1 - \sigma\left(d_j\right)\right)^{1-h_j} \tag{2}$$

where $\sigma(z) = 1/(1 + exp(-z))$and d_j is the deviation of the variable h_j.

$$P(x_i \mid \boldsymbol{h}) = \sigma\left(w_i^T \boldsymbol{h} + b_i\right)^{x_i} \left(1 - \sigma\left(w_i^T \boldsymbol{h} + b_i\right)\right)^{1-x_i} \tag{3}$$

where w_i represents the weight linking the hidden node $\boldsymbol{h}$ with the visible node x_i, while b_i denotes the bias value of the visible node x_i. By integrating Equations (2) and (3), Equation (1) can be derived as Equation (4):

$$\begin{aligned} & P_{\Theta_{LRBN}}(\boldsymbol{x}, \boldsymbol{h}) \\ & = \prod_j \frac{\exp\left(d_j h_j\right)}{1 + \exp\left(d_j\right)} \prod_i \frac{\exp\left(\left(w_i^T \boldsymbol{h} + b_i\right) x_i\right)}{1 + \exp\left(w_i^T \boldsymbol{h} + \boldsymbol{b}\right)} \\ & = \frac{\exp\left(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{x}, \boldsymbol{h})\right)}{\prod_j \left(1 + \exp\left(d_j\right)\right)} \end{aligned} \tag{4}$$

where $\Theta_{LRBN} = \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{d}$, and

$$\begin{aligned} \Gamma_{\Theta_{LRBN}}(\boldsymbol{x}, \boldsymbol{h}) = & -\sum_i \left(w_i^T \boldsymbol{h} + b_i\right) x_i - \sum_j d_j h_j \\ & + \sum_i \log\left(1 + \exp\left(w_i^T \boldsymbol{h} + b_i\right)\right). \end{aligned} \tag{5}$$

Equation (5) bears resemblance to the energy function found in a restricted Boltzmann machine (RBM). However, Equation (5) distinguishes itself with an additional term appended at the end, $\sum_i \log\left(1 + \exp\left(w_i^T \boldsymbol{h} + b_i\right)\right)$. This supplementary term effectively captures the intricate dependencies among the latent variables. Unlike RBMs, where latent

and visible nodes are connected through undirected line segments, the LRBN employs directed connections, resulting in more nuanced interconnections between the latent layers. Consequently, the LRBN possesses superior capability in capturing inherent patterns within the input data compared to RBMs. Moreover, as the LRBN joint distribution is derived from the multiplication of prior and conditional probabilities, it remains unaffected by the collocation function problem encountered in RBM.

3.2. Speech Representation Learning Using LRBN

LRBN is a deep directed generative model designed to learn speech representation from data in an unsupervised manner. Traditional probabilistic deep generative models, such as Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN), often overlook the correlations between latent variables, resulting in a loss of representational power. In contrast, LRBN is capable of capturing complex dependencies between variables by incorporating directed connections between latent and visible layers. To maintain the interdependencies among latent variables, LRBN employs a conditional pseudo-likelihood-based inference method. This approach approximates the true likelihood by considering the conditional probabilities of each visible variable given the values of other visible and latent variables. Notably, visible variables within the LRBN framework can take on continuous or discrete values, while latent variables are binary. Each latent variable is connected to all visible variables, forming a directed acyclic graph.

To utilize LRBN for speech feature representation, it is necessary to preprocess the speech signal to obtain an appropriate input representation. In this particular study, traditional features are employed as input features. The parameters of the LRBN model are learned using the hard Expectation-Maximization (EM) algorithm for training. Notably, the data likelihood is calculated using a max-out method instead of summing out, thereby addressing the challenges associated with traditional EM algorithms in likelihood calculation.

The hard EM algorithm, a variant of the EM algorithm, employs the most probable values of potential variables instead of their expected values to update the model parameters. During the E-step of the hard EM algorithm, Maximum A Posteriori (MAP) inference is required, which can be effectively achieved using a pseudo-likelihood-based approach. To circumvent the exponential number of latent variable configurations, the maximum output setting is utilized in the E-step of learning to approximate the data likelihood. During the M-step, the problem is transformed into parametric learning using complete data, which simplifies the handling process. In this study, the visible layer in LRBN represents the input speech features, while the latent layer represents the learned speech representation. The training objective aims to maximize the log-likelihood of the training data, which is equivalent to minimizing the reconstruction error between the input features and the reconstructed features. Finetuning pretrained models or using them as representation extractors are two common usages [26]. To facilitate comparison with other representation models, we extracted speech representations from a pretrained LRBN model and fine-tuned it on the downstream task, as illustrated in Figure 2.

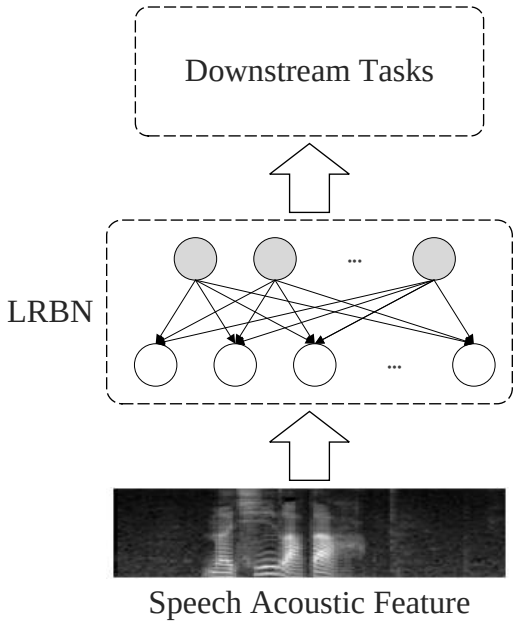

Figure 2. LRBN's speech representations for downstream tasks.

4. Experiment

To evaluate the effectiveness of LRBN in speech representation learning, a series of three experiments were conducted. The first experiment involved a phoneme classification task, wherein the performance of LRBN-extracted speech features was compared with MFCCs, FBank, BNF, Wav2vec 2.0 (W2V2), and Wav2vec 2.0 XLSR (XLSR) features as input features for downstream tasks. This was done to evaluate LRBN's capability to capture high-level speech information. The second experiment focused on a speech recognition task that explored the adaptability of LRBN to five distinct low-resource languages in real-world applications. The performance of LRBN was compared to MFCCs, FBank, BNF, W2V2, and XLSR features as input features for downstream tasks. The third experiment, an ablation experiment, aimed to investigate the effect of varying the number of latent nodes on a single dataset on the performance of LRBN speech representation.

4.1. Datasets

In this study, we evaluated the proposed approach using five low-resource speech datasets, namely the TIBMD Tibetan Lhasa dataset (TI) [27], the Cantonese dataset (CA) in Common Voice [28], the Zeroth-Korean Korean dataset (KO) [29], the THUYG-20 Uyghur dataset (UY) [30], and the IMUT-MC Mongolian dataset (MO) [31]. The TIBMD Tibetan dataset comprised recordings of three dialects (Amdo, Kham, and Ü-Tsang) by native Tibetan-speaking students. For our experiments, we specifically selected 31.8 h of data from the Ü-Tsang dialect. The Cantonese dataset in Common Voice is an open-source speech dataset contributed by volunteers from around the world. We utilized 21.5 h of Cantonese data for our experiments. From the Zeroth-Korean dataset, an open-source dataset for Korean, we randomly selected 26.4 h of data. The THUYG-20 dataset is a Uyghur dataset created by Xinjiang University and Tsinghua University. For our experiments, we utilized 24.9 h of Uyghur data. The IMUT-MC Mongolian dataset consisted of five subsets, with the IMUT-MC3 subset primarily selected as the experimental data. This subset contains daily conversations in Mongolian. All speech signals were sampled at 16 kHz and quantized at 16 bits. Further details regarding each dataset in the continuous speech recognition experiments can be found in Table 1.

Table 1. Data statistics for the speech recognition experiments.

Language	Training Data (h)	Training Utterances	Testing Data (h)	Testing Utterances
TI	28.7	25,704	3.1	2847
CA	19.4	16,884	2.1	1877
KO	23.8	10,224	2.6	1136
UY	21.5	8521	2.4	947
MO	21.8	12,075	3.0	1643

For the phoneme classification experiments, we randomly selected a portion of speech data from each of the aforementioned five datasets. Phoneme-forced alignment was performed using the Kaldi-based Montreal Forced Aligner (MFA). Subsequently, the speech data for each language were segmented into small phoneme-level segments. For specific information regarding the data used for phoneme classification, refer to Table 2.

Table 2. Data statistics for the phoneme classification experiments.

Language	Phoneme Class	Training Data	Testing Data
TI	41	457,881	40,688
CA	71	421,865	37,461
KO	46	586,958	52,166
UY	40	531,711	47,263
MO	32	396,309	35,219

4.2. Experimental Setup

4.2.1. Experimental Setup for Phoneme Classification

In order to utilize LRBN for speech feature representation, a preprocessing step is required to obtain a suitable input representation. We employed MFCCs with a frame length of 25 ms and a frame shift of 10 ms as the input features. The LRBN architecture consists of a visible layer with 39 nodes and a latent layer with 120 nodes. During training, a batch size of 128, a learning rate of 0.00005, and 300 epochs were employed. The training process initializes the parameters randomly, and the activation values of the latent layer are extracted as input features for the phoneme classification task upon completion of the training iterations. For comparision, we also extracted features using MFCCs, FBank, BNF, W2V2, and XLSR. The BNF features were extracted using the Shennong library [32], and the activation values of the bottleneck layer were obtained using the BUT/Phonexia feature extractor [33]. W2V2 features were extracted using the official Facebook-provided wav2vec2-base pre-training model. For the input speech signal, the implicit layer representation of the pre-training model was extracted and weighted, and the weighted representations were summed to obtain the final speech representation. XLSR features utilized the wav2vec2-xlsr pre-training model, and the speech representation extraction followed the same procedure as W2V2.

To ensure a fair comparison, the extracted features were used individually as inputs for the same downstream task, namely the phoneme classification task. We constructed a convolutional neural network model for phoneme classification, consisting of five convolutional layers, two linear layers, and one additional Dropout layer between each layer. The output layer utilized the softmax activation function, with the number of nodes corresponding to the number of phoneme class. The hyperparameters for this model included a learning rate of 0.0008, a batch size of 64, and 100 epochs. The classification accuracy (ACC) was used as the performance metric to evaluate the phoneme classification model.

4.2.2. Experimental Setup for Speech Recognition

Similar to the phoneme classification experiments, the LRBN-based speech feature representation required preprocessing of the speech signal to obtain an appropriate input

representation. In this case, FBank features were chosen, as they are more effective than MFCCs for continuous speech recognition. The FBank features were extracted using a frame length of 25 ms and a frame shift of 10 ms per speech sample. The visible layer of LRBN consisted of 40 nodes, while the latent layer contained 120 nodes. The unsupervised training of LRBN involved a batch size of 32, a learning rate of 0.00001, and 500 epochs. The latent layer activation values were extracted as the input features for the speech recognition task after completing the training iterations. The extraction process for MFCCs, FBank, BNF, W2V2, and XLSR features followed the same procedure as in the phoneme classification experiments.

For the downstream task of speech recognition, we utilized a Transformer model [34]. The convolutional layer of the Transformer comprised two CNN layers with a step size of 2 and a kernel size of 3. The encoder consisted of a stack of six Encoder layers, each containing four attention heads. These attention heads were concatenated and weighted. The hyperparameters for the Transformer model included a learning rate of 0.001, a batch size of 16, and 100 epochs. Batch Normalization was incorporated into the training process to prevent overfitting. The Transformer model was trained for a fixed number of iterations, and the validation set was not used in the experiment. Words were used as the basic modeling unit, and during testing, no language model was employed to decode the sentences. The word error rate (WER) was used as the performance evaluation metric for the model.

4.3. Experimental Results and Analysis

Five distinct phoneme datasets underwent phoneme classification experiments, where LRBN features were tested against traditional features including MFCCs, FBank, and depth model-based BNF. Additionally, W2V2 and XLSR features were utilized as model inputs. The evaluation metric employed was phoneme classification accuracy, and the results can be found in Table 3.

Table 3. Comparison of phoneme classification accuracy for MFCCs, FBank, BNF, W2V2, XLSR, and LRBN features.

Features	ACC(%)				
	TI	**CA**	**KO**	**UY**	**MO**
MFCCs	57.27	43.76	56.43	64.87	62.88
FBank	55.08	42.79	53.48	61.94	62.33
BNF	61.45	55.80	67.13	66.20	67.32
W2V2	85.25	61.56	77.52	71.98	73.77
XLSR	90.10	64.14	86.82	79.07	78.88
LRBN	94.66	64.82 *	87.36 *	79.42 *	80.23

* We evaluated these three results that showed marginal performance improvements for statistically significant differences.

The LRBN features exhibited the highest phoneme classification accuracies across all five languages, achieving percentages of 94.66, 64.82, 87.36, 79.42, and 80.23, respectively. In contrast, traditional methods such as MFCCs and FBank yielded the lowest accuracies, indicating their limited ability to capture higher-level information. Conversely, the supervised learning-based BNF representation method demonstrated a enhancement in phoneme classification accuracy compared to MFCCs and FBank. This observation suggests that the model can effectively capture high-level speech signal information through supervised learning using extensive, well-resourced corpora. Furthermore, the W2V2 and XLSR representations, under the Wav2vec 2.0 self-supervised framework, exhibited superior phoneme classification accuracy when compared to MFCCs, FBank, and BNF representations for all low-resource languages. This finding underscores the potential of self-supervised learning in acquiring phonological representations, especially in low-resource language scenarios. Moreover, the multilingual pre-training model XLSR outperformed the monolingual

pre-training model W2V2 in the phoneme classification task, indicating the benefits of incorporating cross-linguistic information in speech representation learning. Notably, the phoneme recognition accuracy of LRBN surpassed that of XLSR by 4.56, 0.68, 0.54, 0.35, and 1.35 for Tibetan, Cantonese, Korean, Uyghur, and Mongolian, respectively. To evaluate the performance enhancements of our method for Cantonese, Korean, and Uyghur languages, we conducted McNemar's Test [35] to determine whether there are statistically significant differences between LRBN and XLSR. Initially, we stated the null hypothesis that both methods should have the same accuracy rate. Subsequently, for the performance variations of LRBN and XLSR on Cantonese, Korean, and Uyghur, we calculated their respective p-values of 0.000025, 0.00016, and 0.0048. It is widely accepted that a p-value of less than 0.05 indicates a statistically significant difference. In all three datasets, the obtained p-values were less than 0.05, indicating statistically significant differences. This demonstrates the superiority of LRBN in low-resource speech representation, allowing for the capture of complex dependencies between variables through the directed connections between latent and visible layers.

In speech recognition experiments conducted on five low-resource datasets, LRBN, MFCCs, FBank, BNF, W2V2, and XLSR features were employed as model inputs, and their performances were compared using the word error rate as the evaluation metric. The experimental results, displayed in Table 4, reveal that LRBN outperforms all other representations for low-resource languages. Specifically, LRBN achieves the lowest word error rate for Tibetan, Korean, Uyghur, and Mongolian, and the second lowest word error rate for Cantonese. In comparison to traditional representations such as MFCCs and FBank, LRBN demonstrates performance improvements. This indicates that LRBN effectively captures crucial features and speech signal dependencies in low-resource languages, leading to enhanced speech recognition performance. Conversely, the experimental results indicate that BNF exhibits a higher word error rate on the Cantonese dataset compared to traditional features. This suggests that, in supervised learning-based speech representation, the model may fail to capture the necessary acoustic and linguistic information required for low-resource target tasks, particularly when substantial differences exist between the pre-trained source language and the target task language.

Table 4. Comparison of word error rates for MFCCs, FBank, BNF, W2V2, XLSR, and LRBN features.

Features	WER(%)				
	TI	**CA**	**KO**	**UY**	**MO**
MFCCs	37.71	13.94	41.96	45.82	25.21
FBank	36.86	13.58	41.18	44.31	23.63
BNF	36.77	19.93	29.91	43.29	22.13
W2V2	33.68	17.26	28.50	44.18	22.07
XLSR	28.56	10.92	25.75	37.88	20.91
LRBN	27.16	12.53	24.09	36.45	19.78

Moreover, the Wav2vec 2.0 Self-Supervised Framework for W2V2 and XLSR representation methods yielded lower word error rates than traditional and BNF methods for all low-resource languages. However, the high computational cost associated with this framework restricts its practical application. Additionally, the word error rate of W2V2 on the Cantonese dataset exceeded that of the traditional feature FBank by 3.68. This exemplifies the unsuitability of W2V2 for all low-resource languages. In contrast, LRBN represents a lightweight model capable of unsupervised learning, effectively capturing data distribution and latent features while achieving comparable performance to XLSR. The lightweight design and unsupervised learning approach of LRBN make it well-suited for learning speech representations of low-resource languages, resulting in improved performance in low-resource speech recognition tasks.

To mitigate network complexity and overfitting concerns, we conducted ablation studies to examine the impact of the potential number of nodes on LRBN speech represen-

tation performance. Specifically, we performed ablation experiments on the Tibetan speech dataset, as detailed in Table 5.

Table 5. Word error rate comparison for LRBN with varying latent nodes.

Latent Nodes	WER(%)
80	33.20
100	29.47
120	27.16
140	28.73
160	28.36
180	29.09

To gain deeper insights into the results, we also employed a visualization technique, as depicted in Figure 3. From Figure 3, we observe that the speech recognition model achieved optimal performance for the Tibetan language dataset when the number of latent nodes was set to 120. Consequently, in subsequent speech recognition experiments, we fixed the number of latent nodes at 120.

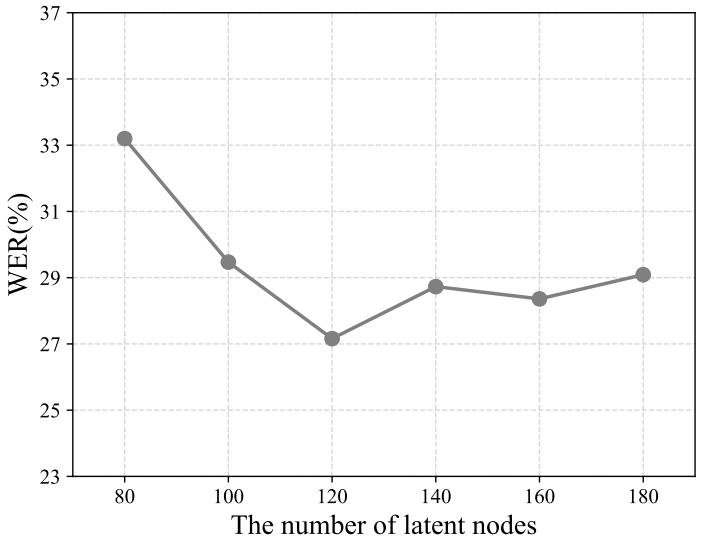

Figure 3. Word error rate comparison for LRBN with varying latent nodes.

5. Conclusions

This paper investigates the effectiveness of LRBN for learning low-resource language speech representations, and compares it to traditional speech representation techniques, namely MFCCs/FBank, supervised speech representation methods such as BNF, and self-supervised speech representation methods such as Wav2vec 2.0. Our experimental results show that LRBN outperforms the other three methods in terms of phoneme classification accuracy and word error rates on five low-resource language speech datasets. Therefore, LRBN is capable of capturing high-level features of speech signals, making it an effective speech representation method for low-resource languages.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z.; software, L.X.; validation, Y.Z., Y.L. and L.X.; formal analysis, Y.Z.; investigation, Y.Z. and L.X.; resources, Y.Z.; data curation, L.X., Y.Z. and Y.L.; writing—original draft preparation, L.X.; writing—review and editing, Y.Z. and Q.J.; visualization, L.X.; supervision, Y.Z.; project administration, Y.Z. and X.X.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61976236.

Data Availability Statement: This paper utilized several datasets, including the TIBMD Tibetan dataset, the Cantonese dataset, the Zeroth-Korean Korean dataset, the THUYG-20 Uyghur dataset, and the IMUT-MC Mongolian dataset. These datasets can be accessed through the following links: http://www.openslr.org/124/ (accessed on 9 October 2022), https://commonvoice.mozilla.org/en/datasets (accessed on 21 February 2023), http://www.openslr.org/40/ (accessed on 21 February 2023), http://www.openslr.org/22/ (accessed on 21 February 2023), http://www.csdata.org/p/687/ (accessed on 21 February 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Veselỳ, K.; Karafiát, M.; Grézl, F.; Janda, M.; Egorova, E. The language-independent bottleneck features. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012 ; IEEE: Piscataway, NJ, USA, 2012, pp. 336–341.
2. Thomas, S.; Seltzer, M.L.; Church, K.; Hermansky, H. Deep neural network features and semi-supervised training for low resource speech recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 6704–6708.
3. Lal, P.; King, S. Cross-lingual automatic speech recognition using tandem features. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2506–2515. [CrossRef]
4. Chung, Y.A.; Hsu, W.N.; Tang, H.; Glass, J. An Unsupervised Autoregressive Model for Speech Representation Learning. In Proceedings of the 20th Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 146–150.
5. Liu, A.T.; Yang, S.W.; Chi, P.H.; Hsu, P.C.; Lee, H.Y. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain , 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6419–6423.
6. Chorowski, J.; Weiss, R.J.; Bengio, S.; Van Den Oord, A. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2041–2053. [CrossRef]
7. Liu, A.T.; Li, S.W.; Lee, H.y. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2351–2366. [CrossRef]
8. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
9. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
10. Zhao, J.; Zhang, W.Q. Improving Automatic Speech Recognition Performance for Low-Resource Languages with Self-Supervised Models. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1227–1241. [CrossRef]
11. Nie, S.; Zhao, Y.; Ji, Q. Latent regression Bayesian network for data representation. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: Piscataway, NJ, USA , 2016; pp. 3494–3499.
12. Nie, S.; Zheng, M.; Ji, Q. The deep regression bayesian network and its applications: Probabilistic deep learning for computer vision. *IEEE Signal Process. Mag.* **2018**, *35*, 101–111. [CrossRef]
13. Tiwari, V. MFCC and its applications in speaker recognition. *Int. J. Emerg. Technol.* **2010**, *1*, 19–22.
14. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of deep representation learning for speech emotion recognition. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1634–1654. [CrossRef]
15. Zhong, G.; Ling, X.; Wang, L.N. From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1255. [CrossRef]
16. Padhi, T.; Biswas, A.; De Wet, F.; van der Westhuizen, E.; Niesler, T. Multilingual bottleneck features for improving ASR performance of code-switched speech in under-resourced languages. *arXiv* **2020**, arXiv:2011.03118.
17. Hermann, E.; Kamper, H.; Goldwater, S. Multilingual and unsupervised subword modeling for zero-resource languages. *Comput. Speech Lang.* **2021**, *65*, 101098. [CrossRef]
18. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proceedings of the 20th Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 3465–3469.
19. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv* **2019**, arXiv:1910.05453.
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
21. Qi, J.; Du, J.; Siniscalchi, S.M.; Lee, C.H. A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1932–1943. [CrossRef]
22. Qi, J.; Du, J.; Siniscalchi, S.M.; Ma, X.; Lee, C.H. Analyzing upper bounds on mean absolute errors for deep neural network-based vector-to-vector regression. *IEEE Trans. Signal Process.* **2020**, *68*, 3411–3422. [CrossRef]

23. Yang, C.H.H.; Qi, J.; Chen, S.Y.C.; Chen, P.Y.; Siniscalchi, S.M.; Ma, X.; Lee, C.H. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6523–6527.
24. Yi, C.; Wang, J.; Cheng, N.; Zhou, S.; Xu, B. Transfer ability of monolingual wav2vec2. 0 for low-resource speech recognition. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
25. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv* **2020**, arXiv:2006.13979.
26. Yang, S.w.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. Superb: Speech processing universal performance benchmark. *arXiv* **2021**, arXiv:2105.01051.
27. Zhao, Y.; Xu, X.; Yue, J.; Song, W.; Li, X.; Wu, L.; Ji, Q. An open speech resource for Tibetan multi-dialect and multitask recognition. *Int. J. Comput. Sci. Eng.* **2020**, *22*, 297–304. [CrossRef]
28. Mozilla. Mozilla Common Voice. Available online: https://commonvoice.mozilla.org (accessed on 21 February 2023).
29. Zeroth Project. Zeroth-Korean: Korean Speech Recognition Corpus for Zeroth ASR. Available online: https://www.openslr.org/40/ (accessed on 21 February 2023).
30. Rouzi, A.; Shi, Y.; Zhang, Z.; Dong, W.; Hamdulla, A.; Fang, Z. THUYG-20: A free Uyghur speech database. *J. Tsinghua Univ. Sci. Technol.* **2017**, *57*, 182–187.
31. Liu, Z.; Ma, Z.; Zhang, X.; Bao, C.; Xie, X.; Zhu, F. IMUT-MC: A speech corpus for Mongolian speech recognition. *China Sci. Data* **2022**, *7*, 13.
32. Bernard, M.; Poli, M.; Karadayi, J.; Dupoux, E. Shennong: A Python toolbox for audio speech features extraction. *Behav. Res. Methods* **2023**, *55*, 1–13. [CrossRef]
33. Silnova, A.; Matejka, P.; Glembek, O.; Plchot, O.; Novotnỳ, O.; Grezl, F.; Schwarz, P.; Burget, L.; Cernockỳ, J. BUT/Phonexia Bottleneck Feature Extractor. In Proceedings of the Odyssey, Les Sables d'Olonne, France, 26–29 June 2018; pp. 283–287.
34. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5884–5888.
35. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef]