Speech Representations and Phoneme Classification for Preserving the Endangered Language of Ladin

Zane Durante[*,1], *Leena Mathur*[*,1,2], *Eric Ye*[1,2], *Sichong Zhao*[1,2], *Tejas Ramdas*[1], *Khalil Iskarous*[2]

[1]Center for Artificial Intelligence in Society [2]Department of Linguistics
University of Southern California, Los Angeles, USA

{durante, lmathur, eye, sichongz, tramdas, kiskarou} @ usc.edu

Abstract

A vast majority of the world's 7,000 spoken languages are predicted to become extinct within this century, including the endangered language of *Ladin* from the Italian Alps. Linguists who work to preserve a language's phonetic and phonological structure can spend hours transcribing each minute of speech from native speakers. To address this problem in the context of Ladin, our paper presents the first analysis of speech representations and machine learning models for classifying 32 phonemes of Ladin. We experimented with a novel dataset of the Fascian dialect of Ladin, collected from native speakers in Italy. We created *frame-level* and *segment-level* speech feature extraction approaches and conducted extensive experiments with 8 different classifiers trained on 9 different speech representations. Our speech representations ranged from traditional features (MFCC, LPC) to features learned with deep neural network models (autoencoders, LSTM autoencoders, and WaveNet). Our highest performing classifier, trained on MFCC representations of speech signals, achieved an 86% average accuracy across all Ladin phonemes. We also obtained average accuracies above 77% for all Ladin phoneme subgroups examined. Our findings contribute insights for learning discriminative Ladin phoneme representations and demonstrate the potential for leveraging machine learning and speech signal processing to preserve Ladin and other endangered languages.

Index Terms: speech representations, phoneme classifiers, machine learning, language preservation, speech signal processing

1. Introduction

Linguists estimate that 60% to 90% of the world's 7,000 spoken languages are likely to become extinct within the next century [1]. The extinction of any language represents an *irreversible* loss of information for humanity, as unique linguistic, psychological, and sociocultural information are embedded within the phonology, syntax, and semantics of each language [2]. The pressing nature of this problem has motivated linguists to conduct research and community interventions to preserve endangered languages [1], [3]. Linguists begin language preservation efforts by collecting spoken data from native speakers and transcribing *phonemes*, the fundamental speech sounds and linguistic units of a language [4], [5]. Since each minute of speech can take up to 2 hours for a trained linguist to transcribe [6], and each hour can take around 100 hours to transcribe [7], manual phoneme classification represents a significant bottleneck in endangered language preservation, motivating the development of computational approaches to assist humans in this task.

Advances in machine learning and speech signal processing have demonstrated the effectiveness of deep neural network models for learning discriminative speech representations [8] and classifying phonemes [9], when trained on large speech datasets. However, existing methods for automated phoneme classification have been largely ineffective for endangered languages [6], due to the limited number of speakers and the difficulty in collecting labeled, transcribed training data. Our paper contributes to endangered language preservation efforts in the specific context of *Ladin*, an endangered language from the Italian Alps. We use a novel dataset of Ladin collected during linguistic field research in Italy with native Ladin speakers. The Ladin people live in multiple valleys around the *Sella* massif mountain range, and the speakers across valleys exhibit a considerable amount of linguistic variation: this research focuses on classifying phonemes in the *Fascian* dialect of Ladin, spoken in the Fassa Valley [10].

To the best of our knowledge, this paper presents the first attempt to develop computational models that classify Ladin phonemes: we contribute a novel analysis of speech representations and machine learning models for classifying 32 phonemes of Ladin [10]. Given the extremely small amount of labeled phoneme data available for endangered languages, including Ladin, our research was driven by the following two questions: (1) Will classical machine learning models outperform deep neural networks when classifying the phonemes of Ladin? (2) Should phoneme classifiers for Ladin be trained on traditional speech representations (e.g., MFCC, LPC) or representations learned by deep unsupervised models, such as autoencoders?

To address these questions, we experimented with 7 classical machine learning models and 1 dense neural network trained on 9 different speech representations that were extracted with traditional algorithms (MFCC, LPC) and autoencoder-based approaches (autoencoder, LSTM autoencoder, WaveNet). Our highest performing model (dense neural network with MFCC representations) achieved an 86% accuracy across all phonemes, accuracies above 90% on 6 individual phonemes, and accuracies above 77% across Ladin phoneme subgroups.

For the endangered language of Ladin, this paper makes two major contributions towards computational approaches that assist linguists in endangered language preservation: (1) We present the first automated phoneme classification approach for Ladin. (2) We present an extensive analysis of speech feature extraction algorithms, speech representations, and phoneme classification models for Ladin. Our overall multi-stage research initiative to preserve Ladin is conceptualized in **Figure 1**.

2. Background

2.1. Endangered language speech research

Phoneme classification has largely relied on classical machine learning models (e.g., Support Vector Machines) trained on acous-

[*]equal contribution, alphabetical order

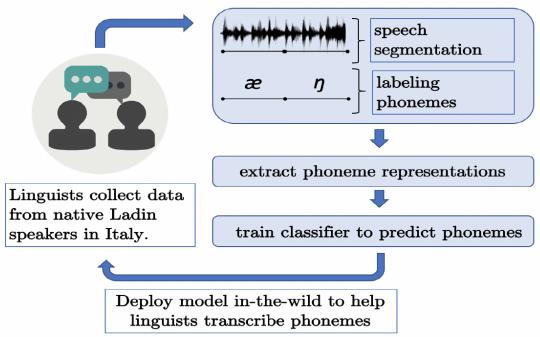

Figure 1: *Multi-stage research process to preserve the Ladin language. This paper focuses on the phoneme representation and classification phases.*

tic waveform representations [11], recurrent neural networks (e.g., LSTM) [9], [12], and convolutional neural networks (e.g., WaveNet) [13]. However, a majority of prior approaches have assumed the existence of hundreds of hours of training data, which is infeasible to collect for low-resource, endangered languages that sometimes only have a few minutes of labeled data available, if at all [6], [14]. Researchers have developed several approaches to address this challenge [15], [16]. Speech recognition systems for low-resource languages have shown improvement through data augmentation [17], [18] and pre-training on high-resource languages [19]. Multilingual models sharing cross-lingual parameters have shown potential in classifying phonemes of the endangered languages of Inuktitut and Tusom [20]. A recent transcription tool for linguists, *Persephone* [21], leveraged recurrent networks trained on limited amounts of the endangered languages of Na (224 minutes) and Chatino (50 minutes).

2.2. Speech representations

A variety of approaches for extracting discriminative representations of speech waveforms have been developed from decades of speech signal processing research [22]. Mel-frequency cepstral coefficients (MFCC), based on cepstral information and the Mel frequency scale, are one of the most commonly used speech representations [23]. Linear Predictive Coding (LPC) approximates formants by applying linear prediction on the spectral domain [22]. More recently, deep neural networks have demonstrated considerable potential for learning useful speech representations [13], [24]. WaveNet, a deep generative model [25], has been leveraged to learn discriminative representations of music and speech [13], [26].

3. Methodology

Our approach involved the following steps: (1) collecting data from native Ladin speakers, (2) extracting traditional speech representations, (3) extracting autoencoder speech representations, and (4) experimenting with machine learning models that were trained on both types of representations to classify Ladin phonemes.

3.1. Ladin dataset collection from native speakers

Our Ladin dataset was collected through 3 months of field research with native Ladin speakers in the Fassa Valley of northern Italy. Our data consist of 76 audio files that span a total of 5 hours and 38 minutes of conversations with four native speakers of Ladin, (three female, one male). The audio was recorded at 44,100 Hz. Teams of linguists trained in the Ladin language were recruited to transcribe the phonemes of two interviews (one female speaker and one male speaker) with a combined duration of 7 minutes and 55 seconds. There are a total of 2520 labeled phoneme segments across both labeled interviews, with each label belonging to one of 33 phoneme classes: 10 vowel classes (39% of the audio), 22 consonant classes (47% of the audio), and 1 class representing the absence of a phoneme (14% silent segments). Vowels spanned two categories with the following distributions: rounded (8%) and unrounded (31%). Consonants spanned 6 key categories with the following distributions: affricates (2%), approximants (8%), fricatives (8%), nasals (7%), plosives (17%), and trills (6%).

3.2. Traditional speech representations

To represent raw speech, we extracted speech features with the traditional algorithms of MFCC and LPC [27]. Given a window size of 25ms that begins at the start of a phoneme, we experimented with two feature extraction approaches, detailed below, for computing speech representations from this window of audio: (1) frame-level extraction and (2) segment-level extraction. **Figure 2** visualizes the relationship between the audio windows, frames, and segments.

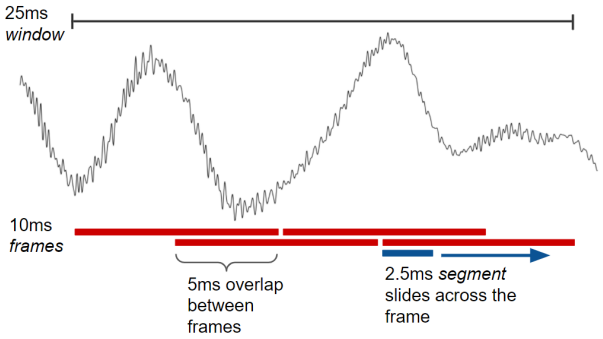

Figure 2: *The 25ms audio window is broken up into four 10ms audio frames, each overlapping by 5ms. For the frame-level approach, features are extracted from each frame. For the segment-level approach, each frame is evaluated by sliding a 2.5ms segment across the frame to calculate the average segment-level features per frame. This sample audio is taken from our dataset, and the line segments are drawn to scale.*

3.2.1. *Frame-level extraction*

The *frame-level* extraction approach divided the 25ms audio window into 4 smaller frames of size 10ms, overlapping each other by 5ms. Features for each 10ms frame were then obtained using MFCC and LPC. Next, features obtained from each frame were concatenated to obtain the *frame-level* representation of the 25ms audio window. For each frame, 12 coefficients were extracted for MFCC, and 8 coefficients were extracted for LPC.

3.2.2. *Segment-level extraction*

The *segment-level* extraction approach used the same 10ms audio frames as the *frame-level* approach. However, for each 10ms frame, features were calculated using a smaller 2.5ms

Table 1: *Average classification accuracy computed across all Ladin phonemes for each of 8 classifiers trained on 9 different traditional and autoencoder speech representations. The best results for each classifier are bolded, and * denotes the highest classification accuracy out of all experiments.*

Representation	Dense NN	SVM (linear)	SVM (rbf)	RF	DT	LR (L1)	LR (L2)	LR (ElasticNet)
Traditional								
MFCC Frame	0.82	0.29	0.43	0.51	0.46	0.46	0.46	0.47
MFCC Segment	**0.86***	0.36	0.52	**0.58**	0.52	0.52	0.53	**0.56**
LPC Frame	0.67	0.23	0.36	0.37	0.35	0.35	0.35	0.35
LPC Segment	0.31	0.15	0.18	0.19	0.18	0.18	0.18	0.19
Autoencoder								
Small AE	0.54	0.28	0.39	0.33	0.37	0.36	0.33	0.38
Big AE	0.50	0.25	0.34	0.31	0.34	0.34	0.32	0.36
Small LSTM AE	0.53	0.28	0.39	0.33	0.37	0.36	0.33	0.38
Big LSTM AE	0.35	0.10	0.16	0.18	0.17	0.17	0.17	0.17
WaveNet	0.77	**0.39**	**0.59**	0.52	**0.56**	**0.54**	**0.57**	0.52

(110 bits) segment that slides across the frame a single bit at a time, creating segments that overlap by 109 bits. Features for each 2.5ms segment were computed and averaged across the frame. Then, the average segment-level features for each frame were concatenated to form the *segment-level* representation. Similar to *frame-level* extraction, 12 coefficients were extracted for MFCC, and 8 coefficients were extracted for LPC. This approach was informed by prior research using Restricted Boltzmann Machines to represent raw speech [28].

3.3. Autoencoder speech representations

Our experiments used traditional autoencoders, LSTM autoencoders, and a WaveNet autoencoder to represent raw speech.

To establish a baseline for how the unlabeled audio files could be leveraged to learn speech representations, we trained two traditional autoencoders (AEs) on the unlabeled Ladin audio, denoted as *Small AE* and *Big AE* with bottleneck layer sizes of 8 and 16, respectively. To explore the possibility of leveraging temporal information encoded in our data, we also trained two LSTM autoencoders [29] on the unlabeled Ladin audio, denoted as *Small LSTM AE* and *Big LSTM AE* with bottleneck layer sizes of 8 and 16, respectively. All models took audio tensor inputs of length 110 (corresponding to 2.5ms of audio data). The initial hidden layer for all models had 2048 neurons, with each subsequent encoding layer having half as many neurons as the previous layer, until the bottleneck layer. The decoding layers mirror each corresponding encoding layer in size. All models used the ReLU activation function on all layers except the final layer, which used tanh, since the audio is bounded on $[-1, 1]$. The bottleneck representations from all autoencoders were extracted per the *segment-level* approach in 3.2.2.

The *WaveNet* extraction approach used a WaveNet autoencoder [26] trained on 306,043 sounds of different pitch, timbre, and envelope, which are common properties of human speech. Our approach was inspired by prior WaveNet autoencoders that learned discriminative representations of English, French, and Mandarin phonemes [13]. The WaveNet autoencoder extracts 16-dimensional encodings from each 512-bit phoneme speech segment. Phoneme segments shorter than 512 were padded with zeros to become valid inputs. Phoneme segments longer than 512 bits were divided into 512-bit segments which were fed into the autoencoder. Encodings were then averaged to obtain the final 16-dimensional speech representation.

3.4. Classification experiments

Phoneme transcription was formulated as a multi-class classification problem to assign each audio segment to one of 33 classes (32 labels for phonemes and 1 label for silent speech segments). To identify effective combinations of speech representations and classification algorithms for Ladin phoneme transcription, we experimented with 8 machine learning classifiers trained on 9 different speech representations. We tested a 2-layer dense neural network (Dense NN, 512 neurons per layer, dropout of 0.05, adam optimizer, 50 epochs), implemented with Keras. We also implemented the following classifiers with the scikit-learn framework: Support Vector Machine (SVM) with a linear and rbf kernel, Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) with L1, L2, and ElasticNet regularization [30]. Due to the small size of our dataset, all experiments were conducted with 5-fold stratified cross-validation, and all models were implemented with default scikit-learn parameters, in order to avoid obtaining overly optimistic model performances [31]. The cross-validation was conducted with a stratified approach to account for imbalanced amounts of data available across phonemes. All features in the training and testing set of each cross-validation fold were standardized per the distribution of the training set. Our primary metric was the phoneme classification accuracy, averaged across folds. A *chance performance* baseline of 3% $(1/33)$ was used to compare our models to a random multi-class classifier.

4. Results and Discussion

4.1. Comparison of Speech Representations

The average classification accuracy computed across all Ladin phonemes for each of 8 classifiers trained on 9 different speech representations is reported in **Table 1**. The highest classifier accuracy obtained was 86%, achieved by a Dense NN trained on the segment-level MFCC speech representation. For each classifier, either the segment-level MFCC or WaveNet speech representations achieved the highest classification accuracy.

For MFCC, the segment-level feature extraction approach always yielded a higher classification accuracy than the frame-level approach. We noted that the efficacy of the frame-level versus segment-level approach depended on the choice of underlying speech processing algorithm used; for LPC, the frame-level approach always yielded a higher classification accuracy than the segment-level approach. The poor performance of the segment-level LPC representation is likely due to the small size

of the segments; we observed greater variance within frames for the segment-level LPC coefficients compared to the segment-level MFCC coefficients.

Representations learned from the traditional autoencoders trained on the unlabeled Ladin speech files were outperformed by the segment-level MFCC and WaveNet speech representations. We noted that Big AE always outperformed Big LSTM AE across all classifiers. This is likely due to the limited amount of training data and the larger number of trainable parameters for LSTM networks. Small AE and Small LSTM AE exhibited similar performance across all classifiers. The small autoencoders always outperformed their larger counterparts, indicating that they learned more useful, condensed representations of phonemes in our dataset's context.

4.2. Comparison of Phoneme Classifiers

We noted that the Dense NN achieved the highest accuracy across all speech representations. These results suggest that deep neural network models for phoneme classification can outperform classical machine learning algorithms, even in low-resource language settings. In contrast, we noted that the linear SVM achieved the lowest accuracy across all speech representations. However, using an RBF kernel substantially increased SVM classification accuracy across all speech representations.

4.3. Phoneme Classification Performances

We evaluated the ability of Dense NN trained on MFCC segment representations to classify individual Ladin phonemes and phoneme subgroups, documented in **Table 2**. All phonemes and phoneme subgroups had classification accuracies substantially higher than our chance level baseline (3%), demonstrating the potential for leveraging discriminative MFCC segment representations to classify Ladin phonemes.

The average vowel classification accuracy (88%) was higher than that of consonants (83%), and average unrounded vowel classification accuracy (90%) was higher than that of rounded vowels (80%). The performance difference between unrounded and rounded vowels may have been influenced by the impact of rounding on speech formants: rounding lengthens the vocal tract, lowering all formants on the spectral envelope [32]. Since MFCC represents spectral envelopes of sounds, features extracted from rounded vowels may have been less discriminative. Within the consonants, fricatives and nasals had the highest average classification accuracy (87%), followed by plosives and trills (82%), affricates (80%) and approximants (77%). Affricates are more temporally complex sounds, as they combine a fricative and plosive, which may contribute to their lower performance. Approximants form during slight constrictions of the vocal tract that are not as prominent or turbulent as constrictions used to produce fricatives; the complex and subtle nature of approximants may contribute to their lower performance [32]. *Our results for Ladin phoneme subgroup classification demonstrate the potential for using MFCC segment-level approaches to compute discriminative representations of Ladin across subgroups of speech sounds*. These findings motivate further development of techniques to improve the classification accuracy of more complex Ladin phoneme groups.

An automated phoneme classification system deployed in-the-wild to help linguists rapidly transcribe Ladin must perform effectively on individual phonemes, in addition to phoneme subgroups. Six individual phonemes exhibited classification accuracies above 90%: the unrounded vowels *i*, *e*, and *a*, the fricatives *ʃ* and *s*, and the nasal *m*. The lowest-performing individual phonemes were the rounded vowel *œ*, the fricative *z*, the unrounded vowel *ɛ*, the plosive *b*, and the nasals *ŋ* and *ɲ*; the low performances of these individual phonemes, compared to the other phonemes in their subgroups, may have resulted from the imbalanced distribution of phonemes in the dataset. *Our results for individual Ladin phoneme classification demonstrate the potential for using MFCC segment-level approaches to learn discriminative representations of individual Ladin speech sounds*.

Table 2: *Average classification accuracy for each Ladin phoneme subgroup, the 6 phonemes with the highest performance, and the 6 phonemes with the lowest performance, obtained by Dense NN trained on MFCC Segment representations.*

Phonemes	Classification Accuracy
Phoneme Subgroups	
All Vowels	0.88
Rounded Vowels	0.80
Unrounded Vowels	0.90
All Consonants	0.83
Affricates	0.80
Approximants	0.77
Fricatives	0.87
Nasals	0.87
Plosives	0.82
Trills	0.82
Highest Phoneme Performances	
i	0.95
e	0.93
ʃ	0.93
a	0.90
s	0.90
m	0.90
Lowest Phoneme Performances	
œ	0.50
z	0.62
ɛ	0.63
b	0.67
ŋ	0.67
ɱ	0.67

5. Conclusions

Our paper presents the first analysis of speech representations and machine learning models for classifying 32 phonemes of Ladin. We demonstrate the potential for learning discriminative representations of Ladin phonemes through traditional speech processing algorithms (MFCC in particular) and autoencoder-based approaches. We provide a proof-of-concept to recruit linguists to further transcribe recordings in our novel Ladin dataset.

Future research includes experimenting with data augmentation [18] to increase the quantity of Ladin training data and developing multilingual models that learn cross-lingual phoneme representations to improve Ladin phoneme classifiers. Ladin shares phonetic and phonological attributes with non-endangered languages (e.g., Italian) [10], indicating that cross-lingual modeling may be a promising research direction for Ladin preservation. Future work also includes integrating our machine learning models into a real-world system to help linguists rapidly transcribe Ladin during field research (Figure 1). Our work contributes insights for the future development and deployment of phoneme classification approaches that can help linguists preserve Ladin and other low-resource endangered languages.

6. Acknowledgements

We thank the native speakers of Ladin for their generous participation in this research and contribution of recordings. We thank the Ladin Cultural Institute in Italy and the University of Southern California for supporting this research.

7. References

[1] S. Romaine, "Preserving endangered languages," *Language and Linguistics Compass*, pp. 115–132, 2007.

[2] K. Hale, M. Krauss, L. J. Watahomigie, *et al.*, "Endangered languages," *Language*, vol. 68, no. 1, pp. 1–42, 1992, ISSN: 00978507, 15350665.

[3] L. A. Grenoble and L. J. Whaley, *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press, 2005, ISBN: 9780521816212.

[4] S. Chapman and C. Routledge, *Key Ideas in Linguistics and the Philosophy of Language*. Edinburgh University Press, 2009, ISBN: 9780748626182.

[5] P. Bhaskararao, "Phonetic documentation of endangered languages: Creating a knowledge- base containing sound recording, transcription and analysis," *Acoustical Science and Technology*, pp. 219–226, 2004.

[6] O. Adams *et al.*, "Evaluating phonemic transcription of low-resource tonal languages for language documentation," in *Proceedings of LREC*, 2018.

[7] M. Ćavar, D. Ćavar, and H. Cruz, "Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, ASR," in *Proceedings of LREC*, 2016.

[8] S. Latif *et al.*, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *ArXiv*, vol. abs/2001.00378, 2020.

[9] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," in *Proceedings IEEE International Joint Conference on Neural Networks*, vol. 4, 2005, pp. 2047–2052.

[10] Y. Yang, R. Walker, A. Vietti, *et al.*, "Ladin, varieties of val di fassa," *Journal of the International Phonetic Association*, pp. 1–26, 2021.

[11] J. Yousafzai, Z. Cvetkovic, and P. Sollich, "Tuning support vector machines for robust phoneme classification with acoustic waveforms," in *INTERSPEECH*, 2009.

[12] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans Neural Net*, 1994.

[13] J. Chorowski *et al.*, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019, ISSN: 2329-9290.

[14] T. Tsunoda, *Language endangerment and language revitalization: An introduction*. De Gruyter Mouton, 2006.

[15] C. Soria, Ed., *Computational Linguistics for Low-Resource Languages*. MDPI, 2019, Issue of *Information*.

[16] A. Anastasopoulos, "Computational tools for endangered language documentation," Ph.D. dissertation, University of Michigan, 2019.

[17] B. Thai, "Deepfake detection and low-resource language speech recognition using deep learning," M.S. thesis, Rochester Institute of Technology, 2019.

[18] A. Ragni *et al.*, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*, International Speech Communication Association (ISCA), 2014, pp. 810–814.

[19] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7909–7913.

[20] X. Li *et al.*, "Universal phone recognition with a multilingual allophone system," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[21] A. Michaud *et al.*, "Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit," *Language Documentation and Conservation*, vol. 12, 2018.

[22] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," in *From Natural to Artificial Intelligence*, R. Lopez-Ruiz, Ed., Rijeka: IntechOpen, 2018, ch. 1.

[23] A. Lawson *et al.*, "Survey and evaluation of acoustic features for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5444–5447.

[24] Y.-A. Chung *et al.*, "An unsupervised autoregressive model for speech representation learning," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds., ISCA, 2019, pp. 146–150.

[25] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," vol. abs/1609.03499, 2016.

[26] J. Engel *et al.*, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 1068–1077.

[27] B. McFee, V. Lostanlen, A. Metsai, *et al.*, *Librosa/librosa: 0.8.0*, version 0.8.0, Jul. 2020.

[28] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5884–5887.

[29] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, Lille, France: JMLR.org, 2015, pp. 843–852.

[30] L. Buitinck *et al.*, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop*, 2013.

[31] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *ArXiv*, vol. abs/1811.12808, 2018.

[32] E. Zsiga, *The Sounds of Language: An Introduction to Phonetics and Phonology*. Wiley-Blackwell Publishers, 2013.