Analysis of Phoneme Recognition Models

Shreya Pandey (M24CSA030), Princu Singh (M24CSA024)

Abstract

Phoneme recognition is a crucial task in automatic speech recognition (ASR) and related fields, aiming to identify and classify phonemes in speech. Recent advancements in self-supervised learning (SSL) have significantly improved phoneme recognition by leveraging large-scale unlabeled data. This paper presents a comparative analysis of three state-of-the-art SSL models: Wav2Vec 2.0 [1], HuBERT [2], and WavLM [3]. We evaluate these models based on their methodologies, strengths, limitations, and performance on various speech-processing benchmarks. While Wav2Vec 2.0 excels in ASR, HuBERT enhances phoneme representation learning, and WavLM demonstrates superior performance in multi-speaker and noisy environments. We further discuss open challenges and future research opportunities to improve SSL-based phoneme recognition systems.

1 Introduction: The Task and Its Real-World Importance

Phoneme recognition is an essential component of automatic speech recognition (ASR) and related applications such as speaker verification, speech separation, and diarization. Traditional supervised learning approaches rely on large labeled datasets, which are expensive to curate, particularly for low-resource languages.

Self-Supervised Learning (SSL) methods have emerged as a transformative solution, enabling models to learn from vast amounts of unlabeled speech data. State-of-the-art models like **Wav2Vec 2.0** [1], **HuBERT** [2], and **WavLM** [3] enhance phoneme recognition and speech processing. Each model introduces innovations in feature learning, robustness, and generalization.

2 GitHub repository

The complete task can be accessed from this repository. Question1

3 Strengths and Limitations of State-of-the-Art Models

3.1 Wav2Vec 2.0

Method: Utilizes contrastive learning to train raw speech waveforms by distinguishing between positive and negative examples [1].
Strengths:

- **Self-Supervised Pre-Training:** The model allows leveraging huge amounts of unlabeled speech data through self-supervised pre-training, enabling it to acquire strong speech representations without the need for transcribed data. This is a major advantage over purely supervised approaches.

- **Low-Resource Performance:** One of the key strengths is the model's ability to achieve solid performance with very limited labeled data:
  - Achieves 4.8/8.2 WER on Librispeech test-clean/other with just 10 minutes of labeled data.
  - Outperforms previous state-of-the-art using 1 hour of labeled data with 100 hours of training.

  This highlights the model's efficiency for low-resource scenarios.
- **Scalability:** The model scales effectively with more unlabeled pre-training data and larger model sizes.
  - The LARGE model outperforms the BASE model across all labeled data settings.
  - Pre-training on 60,000 hours of LibriVox data provides gains over the 960-hour Librispeech dataset.
- **State-of-the-Art Results:** Using all 960 hours of labeled Librispeech data, wav2vec 2.0 achieves state-of-the-art performance with 1.8/3.3 WER on test-clean/other.
- **Phoneme Recognition:** The model achieves state-of-the-art results on TIMIT phoneme recognition tasks, demonstrating its ability to learn fine-grained speech units.

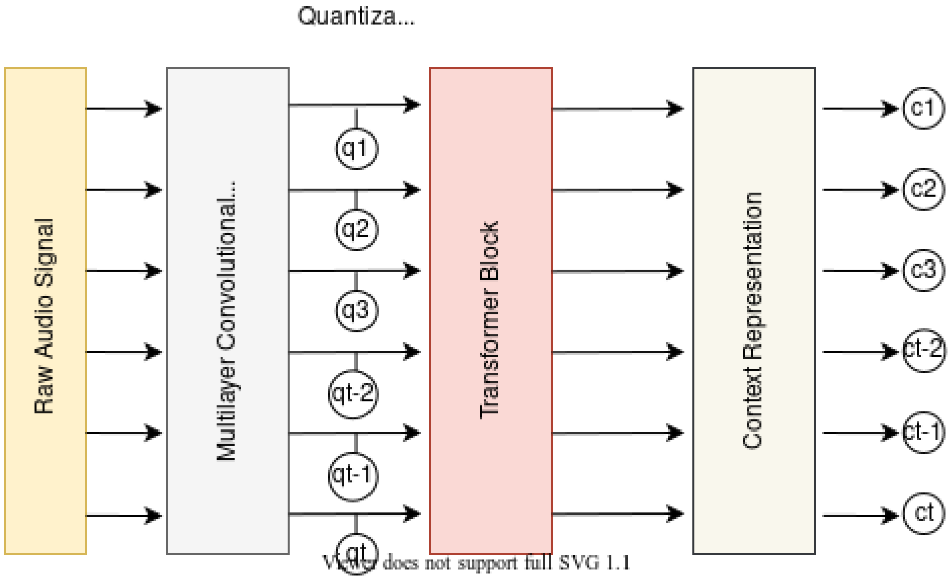


Figure 1: Wav2Vec Explained

Limitations:

- **Required Computing Power:** Pre-training the LARGE model on LibriVox required approximately 128 GPUs for 5.2 days, resulting in a very high computational cost that may be prohibitive for many researchers.

- **Localized Dependence on English:** The model has been primarily evaluated on English datasets (Librispeech, TIMIT). Its effectiveness for other languages, especially low-resource languages, remains uncertain.
- **Decoding Methodology:** The decoding mechanism is relatively simple compared to CTC. More advanced decoding methods, such as attention-based sequence-to-sequence models, could potentially enhance performance.
- **Limitations of Metrics:** The primary metric used, Word Error Rate (WER), has several limitations:
  - It does not account for the severity of errors (e.g., minor misspellings vs. completely incorrect words).
  - It may not fully capture the model's understanding of speech semantics.
- **Language Model Dependency:** While the model performs well independently, its performance improves significantly when combined with language models. However, this adds a layer of complexity and may reduce applicability in certain problem domains.

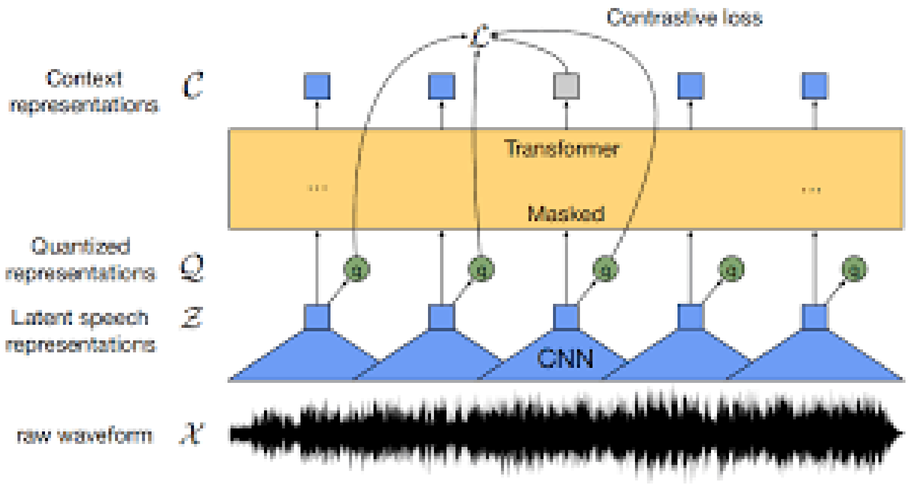


Figure 2: Model Architecture used in Wav2Vec

3.2 HuBERT

Method: Employs masked speech prediction, where clustered hidden units act as training targets [2].
Strengths:

- **Self-supervised learning framework**: HuBERT uses masked prediction of clustered acoustic units, combining acoustic and language modeling:
  - Utilizes iterative k-means clustering on MFCC or learned features for pseudo-labeling.
  - Only masked regions are predicted (loss weight $\alpha = 1$), strongly orienting the model towards inferring context and structure.

  - Demonstrates improved performance over wav2vec 2.0 and DiscreteBERT in low-resource settings (e.g., 4.6% WER on LibriSpeech test-clean with 10 minutes of labeled data).

- **Scalability and Low-Resource Adaptability**:
  - Model sizes: BASE (95M), LARGE (317M), and X-LARGE (1B parameters).
  - With 10 minutes of labeled data, HuBERT X-LARGE achieves 4.6%/6.8% WER (test-clean/other), providing an 18% relative reduction compared to wav2vec 2.0 in one-hour labeled settings.

- **Iterative Cluster Refinement**:
  - Quality of clustering improves over iterations:
  - First iteration (MFCC features): PNMI = 0.2871.
  - Second iteration (features from the 6th transformer layer): PNMI = 0.6861.
  - Enhances phoneme recognition accuracy for improved ASR performance.

- **Efficient Training**:
  - Avoids complications of contrastive loss such as negative sampling and diversity loss.
  - Achieves similar performance to wav2vec 2.0 but with fewer training steps.

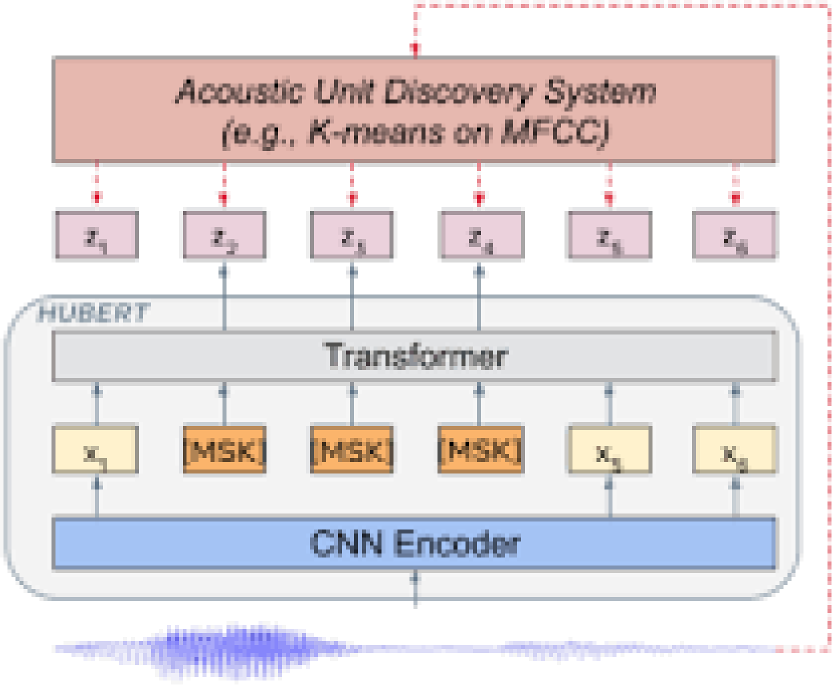


Fig. 1. The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

Figure 3: Model Architecture used in Hubert

Limitations and Challenges:

- **Computational Requirements**:
  - Pretraining HuBERT X-LARGE on 60k hours requires 256 GPUs, limiting accessibility.
  - Iterative clustering and model refinement are time-consuming.
- **Language and Domain Constraints**:
  - Evaluated exclusively on English datasets (LibriSpeech, Libri-light).
  - Multilingual capabilities remain unverified despite the use of mixed-domain data.
- **Metric Limitations**:
  - **Word Error Rate (WER)**:
    - Does not differentiate between minor (e.g., homophones) and critical errors.
    - Best results require external language models (e.g., 23% WER reduction with Transformer LM).
  - **Cluster Purity and PNMI**:
    - Measure clustering quality but do not accurately reflect downstream task performance.
- **Architectural Trade-offs**:
  - Fixed feature encoder limits adaptation to new domains/languages during fine-tuning.
  - Middle layers (6th to 9th) provide the best clustering features, while last layers degrade in quality.
  - Different layers must be optimized for speaker vs. content tasks.
- **Clustering Stability**:
  - K-means clustering on MFCCs achieves a local PNMI of 0.287, while HuBERT features improve this to 0.686.
  - K-means is sensitive to initialization and data subsampling.

Comparative Results and Metrics:

- **Word Error Rate (WER)**:
  - Primary metric for discourse analysis in ASR evaluation.
  - HuBERT X-LARGE recorded 1.8%/2.9% WER on test-clean/other with 960h labeled data.
  - Outperformed Conformer and wav2vec 2.0 in low-resource scenarios.
- **Phone-Normalized Mutual Information (PNMI)**:
  - Measures how well clusters align with ground-truth phonemes.
  - Iterative refinement improved PNMI from 0.287 to 0.686.
- **Cluster Purity**:

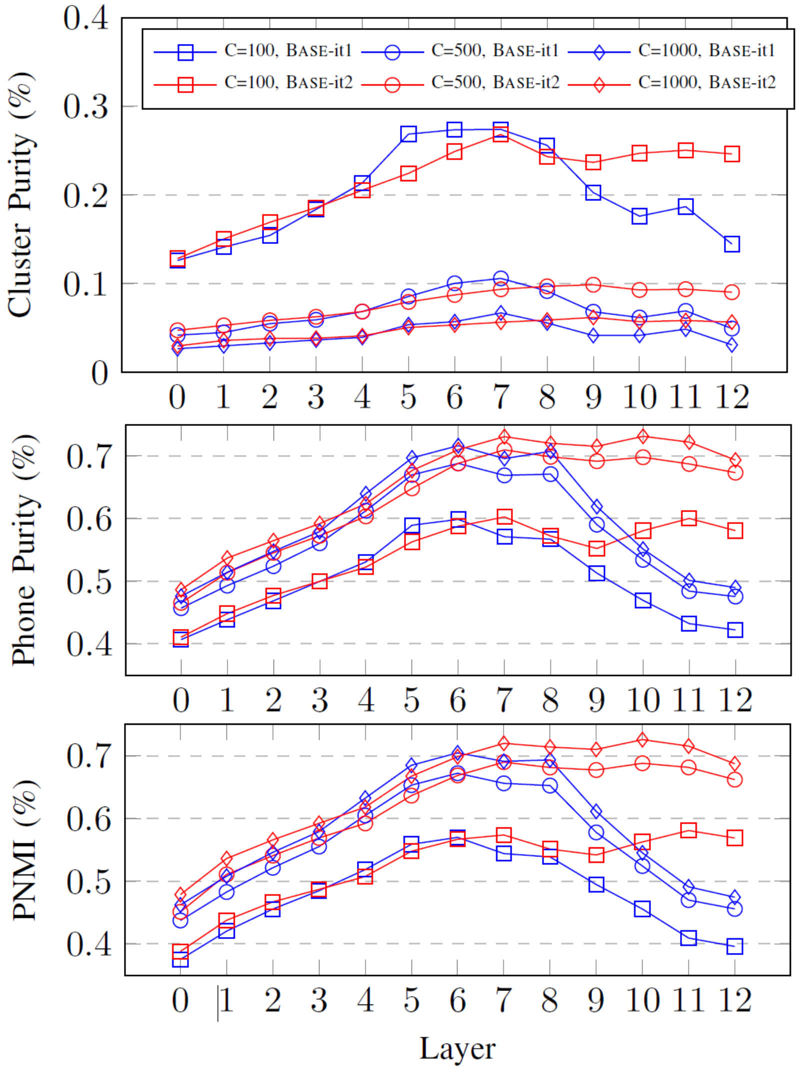


Figure 4: Quality of the cluster assignments obtained by running k-means clustering on features extracted from each transformer layer of the first and the second iteration BASE HuBERT models

– Evaluates cluster assignments; higher values indicate better quality.

Strengths of Metrics:

- WER provides a standard measure to compare the performance of diverse ASR systems.
- PNMI quantifies acoustic-phonetic correctness, essential for iterative refinement.

Limitations of Metrics:

- WER does not consider semantics or the impact of errors on meaning.
- Cluster purity and PNMI lack task-specific performance evaluation, such as speaker verification.

3.3 WavLM

Method: Extends HuBERT by incorporating masked speech denoising, improving multi-speaker and noisy speech recognition [3].
Strengths:

- **Universal Speech Representation Learning**: WavLM achieved state-of-the-art performance across 19 tasks on the SUPERB benchmark, including speaker verification (ASV), speech separation (SS), and automatic speech recognition (ASR). Key results include:
  - **Speaker Verification**: 0.383% EER on VoxCeleb1, a 62% relative reduction over ECAPA-TDNN.
  - **Speech Separation**: 27.7% WER on LibriCSS, improving upon Conformer models.
  - **Speaker Diarization**: 10.35% DER on CALLHOME, achieving a 12.6% improvement over prior methods.
- **Creative Pre-Trained Structure**:
  - **Masked Speech Denoising**: Simulated noisy/overlapped speech during pretraining leads to robust multi-speaker modeling, reducing DER by 22.6% compared to HuBERT.
  - **Gated Relative Position Bias**: Replaces convolutional position embeddings with adaptive gates, improving ASR performance to 3.44% WER on LibriSpeech test-other.
- **Scalability and Data Diversity**:
  - Trained on 94k hours of diverse data (LibriLight, GigaSpeech, VoxPopuli), reducing audiobook bias from prior models like wav2vec 2.0.
  - Large model (316M parameters) achieves 2.1% WER on LibriSpeech with 960 hours of labeled data, setting a new supervised state-of-the-art.
- **Low-Resource Adaptability**: WavLM Base+ achieves 5.4%/9.8% WER with only 1 hour of labeled data (test-clean/other), an 18% improvement over wav2vec 2.0.

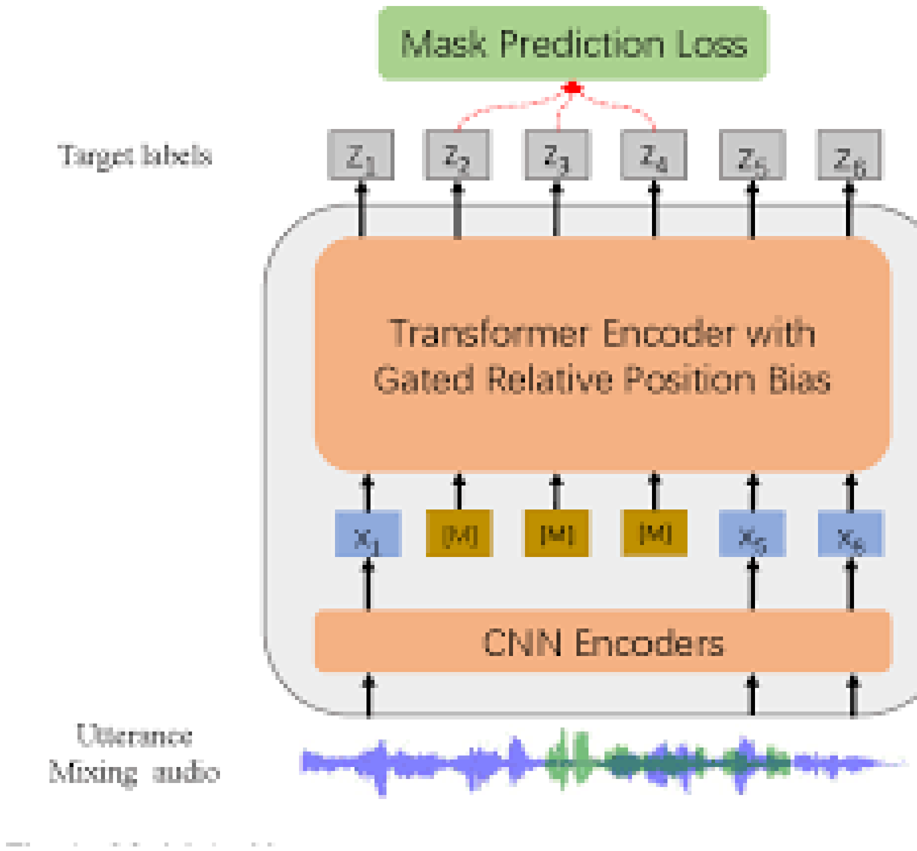
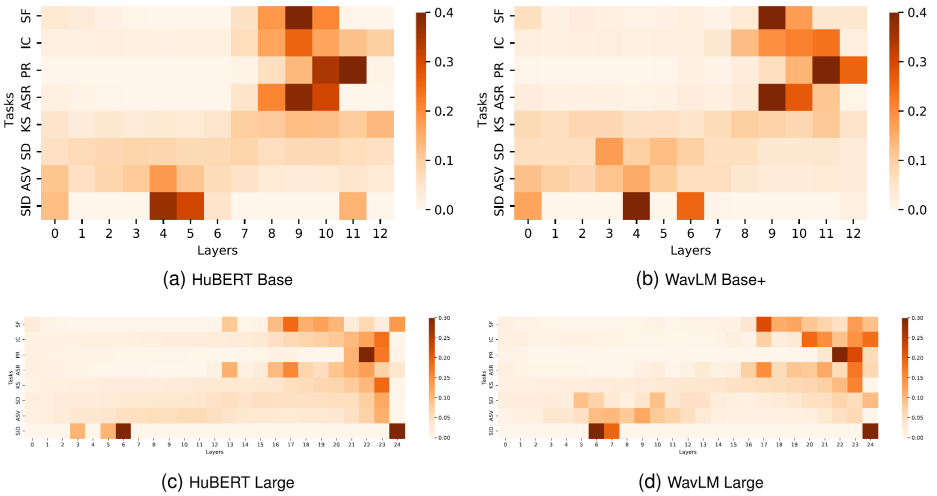


Figure 5: Model Architecture used in WavLM

Limitations:

- **Computation Requirements**: Pre-training requires 64 V100 GPUs for 700k steps, making it inaccessible for smaller research teams.
- **Task-Specific Overfitting**: Unfreezing pre-trained parameters during speech separation fine-tuning increases WER from 6.0% to 8.2%, indicating overfitting to simulated data.
- **Language and Domain Limitations**: Evaluated primarily on English datasets (LibriSpeech, VoxCeleb). Some multilingual verification was done using the Vox-Populi corpus.
- **Metric Limitations**:
  - **Word Error Rate (WER)**:
    - Does not differentiate between minor (e.g., homophones) and major errors.
    - Relies on external language models for optimal performance (e.g., a Transformer LM reduces WER by 23%).
  - **Diarization Error Rate (DER)**: Does not proportionally penalize overlapping speech errors.
  - **Equal Error Rate (EER)**: Sensitive to dataset distribution shifts (e.g., VoxCeleb vs. real-world noisy data).
- **Architecture Trade-offs**:

- Freezing the feature encoder during fine-tuning limits adaptation to low-level acoustic variations.
- Layer-wise analysis reveals that speaker-related tasks rely on middle layers, while content tasks depend on top layers, making multi-task optimization challenging.



(a) HuBERT Base

(b) WavLM Base+

(c) HuBERT Large

(d) WavLM Large

Figure 6: Weight analysis on the SUPERB Benchmark

4 Conclusion

Wav2Vec 2.0 excels in ASR, HuBERT improves phoneme learning, and WavLM is the most robust for multi-speaker speech tasks. Future research should focus on **low-resource adaptation, computational efficiency, and real-world robustness**.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-S. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.

[3] S. Chen, Y. Wu, J. Gao, Z. Yao, C. Zhang, S. Liu, H. Meng, and M. Zhou, “Wavlm: Large-scale self-supervised pre-training for full-stack speech processing,” IEEE Journal of Selected Topics in Signal Processing, 2022.

[4] A. Baevski, M. Auli, and A. Mohamed, “Effectiveness of self-supervised pre-training for speech recognition,” arXiv, abs/1911.03912, 2019.

[5] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, “Improving transformer-based speech recognition using unsupervised pre-training,” arXiv, abs/1910.09932, 2019.

[6] W. Chan, D.S. Park, C.A. Lee, Y. Zhang, Q. Le, and M. Norouzi, “Speech-Stew: Simply mix all available speech recognition data to train one large neural network,” arXiv, 2021, arXiv:2104.02133.

[7] S. Horiguchi, S. Watanabe, P. Garcia, Y. Xue, Y. Takashima, and Y. Kawaguchi, “Towards neural diarization for unlimited numbers of speakers using global and local attractors,” IEEE Autom. Speech Recognit. Understanding Workshop, Cartagena, Colombia, pp. 98–105, Dec. 13–17, 2021, doi: 10.1109/ASRU51503.2021.9687875.

[8] W.-N. Hsu, A. Lee, G. Synnaeve, and A. Hannun, “Semisupervised speech recognition via local prior matching,” arXiv preprint arXiv:2002.10336, 2020.

[9] W.-N. Hsu, Y. Zhang, and J. Glass, “Learning latent representations for speech generation and transformation,” in INTERSPEECH, 2017.