The Complementary Role of Para-linguistic cues for Robust Pronunciation Assessment

Yassine El Kheir
DFKI
yassine.el_kheir@dfki.de

Shammur Absar Chowdhury
QCRI
schowdhury@hbku.edu.qa

Ahmed Ali
HUMAIN
ahmed.ali@humain.com

Abstract

Research on pronunciation assessment systems focuses on utilizing phonetic and phonological aspects of non-native (L2) speech, often neglecting the rich layer of information hidden within the para-linguistic cues. In this study, we proposed a novel pronunciation assessment framework, **IntraVerbalPA**. The framework innovatively incorporates both fine-grained frame- and abstract utterance-level para-linguistic cues, alongside the raw speech and phoneme representations. Additionally, we introduce the "Goodness of phonemic-duration" metric to effectively model phoneme duration distribution within the framework. Our results validate the effectiveness of the proposed IntraVerbalPA framework and its individual components, yielding performance that matches or outperforms existing research works.

1 Introduction

Computer-assisted pronunciation training (CAPT) for foreign language learning has seen a surge in global demand in recent years. CAPT benefits non-native learners with personalized, cost-effective feedback, promotes self-directed learning and improves pronunciation skills. It also offers flexibility compared to traditional instruction (Eskenazi, 2009; Litman et al., 2018; Kheir et al., 2023). One of the main objectives of the CAPT is to automate pronunciation assessment (PA). To achieve this goal, the automated PA model needs to estimate a score that reflects oral proficiency based on some standardized assessment criteria (Levy and Stockwell, 2013; Eskenazi, 2009).

The task of PA is inherently subjective, even scores assigned by human expert annotators often vary for the same spoken utterance. These discrepancies arise from annotator's unique experiences, their interpretations of the scoring guidelines, and/or their focus on specific aspects of pronunciation – like fluency, prosody, word accuracy, or even a combination. Hence, designing an automated PA that emulates the annotators' (or a teacher) is very much complex and challenging. The challenges extend beyond the constraints of dataset availability, and modeling intricacies, to include the crucial task of selecting features and approaches to model their representations. Numerous investigations have explored a range of features and modeling approaches to enhance modeling performance. These explorations have encompassed the utilization of Goodness-of-Pronunciation (GOP) metrics (Lin et al., 2020; Gong et al., 2022; Hu et al., 2015), the integration of manually crafted handful of para-linguistic features such as duration, energy, and pitch (Zhang et al., 2021a; Chao et al., 2022; Chen et al., 2023), as well as the utilization of state-of-the-art pre-trained self-supervised learning models for modeling improvement (Kim et al., 2022; Lin and Wang, 2023; Yang et al., 2022). However, the majority of the studies often neglect the rich layer of information hidden within the para-linguistic cues. For automated PA, integrating an additional layer of para-linguistic cues – pitch, intonation, voice quality, etc., can greatly enhance the system's ability to evaluate oral proficiency, bringing the human perception factor into the equation. Therefore, we introduce a novel PA framework **IntraVerbalPA**. The framework is jointly trained to score the 'fluency' and 'prosodic' aspects in multi-task setups. IntraVerbalPA leverages both latent speech and phoneme embedding while complementing them with handcrafted frame- and utterance-level para-linguistic paralinguistic cues.

2 Proposed Framework

Figure 1 shows our proposed IntraVerbalPA framework, designed to train an efficient end-to-end pronunciation assessment model using different sources of information from the input signal. The IntraVerbalPA model comprised of 4 mod-

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 302–306
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

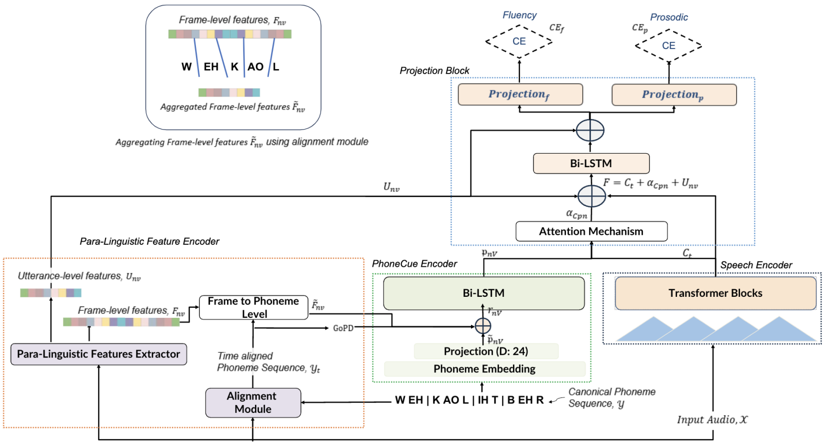
Frame-level features, F_{nv}
W
EH
K
AO
L
Aggregated Frame-level features $\tilde{F}_{nv}$
Aggregating Frame-level features $\tilde{F}_{nv}$ using alignment module
Fluency
CE_f
CE
Prosodic
CE_p
CE
Projection Block
$Projection_f$
$Projection_p$
Bi-LSTM
U_{nv}
$F = C_t + \alpha_{Cpn} + U_{nv}$
α_{Cpn}
Attention Mechanism
Para-Linguistic Feature Encoder
p_{nv}
C_t
PhoneCue Encoder
Speech Encoder
Utterance-level features, U_{nv}
Frame-level features, F_{nv}
Frame to Phoneme Level
$\tilde{F}_{nv}$
Bi-LSTM
r_{nv}
Transformer Blocks
GoPD
$\tilde{p}_{nv}$
Para-Linguistic Features Extractor
Projection (D: 24)
Phoneme Embedding
Time aligned Phoneme Sequence, y_t
Alignment Module
W EH | K AO L | IH T | B EH R
Canonical Phoneme Sequence, y
Input Audio, X

Figure 1: Overview of proposed IntraVerbalPA.

ules, *Speech Encoder*, *PhoneCue Encoder*, *para-linguistic Features Encoder*, and a *Projection Block*.

Framework Overview Given an input raw signal $\mathcal{X}$, of n samples, we first extract contextualized acoustic representations, C_t (of dimension, D : 1024), from the **Speech Encoder**. Simultaneously, $\mathcal{X}$ is also passed through the **para-linguistic Features Encoder** to obtain para-linguistic phoneme-level ($\tilde{\mathbf{F}}_{\mathbf{nv}}$), utterance-level $\mathbf{U}_{\mathbf{nv}}$ feature along with *duration*, $GoPD$ representation. We then pass $\tilde{\mathbf{F}}_{\mathbf{nv}}$ and $GoPD$ to the **PhoneCue Encoder**. The resultant output, $\tilde{\mathbf{p}}_{\mathbf{nv}}$, along with C_t, and $\mathbf{U}_{\mathbf{nv}}$ are then passed to the **Projection Block** for predicting Fluency and Prosodic scores.

2.1 Speech Encoder Module

The wav2vec2-large (Conneau et al., 2020) model is a pre-trained wav2vec2.0 (Baevski et al., 2020). It follows the same architecture as the wav2vec2.0 model.

2.2 Para-linguistic Features Encoder

Inside the para-linguistic feature encoder, using the input $\mathcal{X}$, we first extract low-level descriptors in frame-level ($\mathbf{F}_{\mathbf{nv}}$) and functionals to create utterance-level ($\mathbf{U}_{\mathbf{nv}}$.) representation using OpenSmile. We then align the input $\mathcal{X}$ with the canonical phoneme sequence $\mathcal{Y}$ using the *Alignment Module* to convert frame-level para-linguistic $\mathbf{F}_{\mathbf{nv}}$ representation to phoneme-level ($\tilde{\mathbf{F}}_{\mathbf{nv}}$) representation. Moreover, we also use the phoneme-level alignments to calculate the *duration representation*, $GoPD$.

2.2.1 Alignment Module

To align the canonical sequence with the audio, we opt for wav2vec2.0 trained for frame-level classification (Zhu et al., 2022).

2.2.2 Goodness of phonemic-duration (GoPD)

We present a novel metric called Goodness of phonemic-duration (GoPD), drawing inspiration from the Goodness of Pronunciation (GoP) metric introduced in (Witt and Young, 2000). The GoP metric is defined for a given observation $\mathbf{O}$ and a phone $\mathbf{p}$ by the following equation:

$$GOP(p) = P(p|O) = \frac{p(O|p)\ P(p)}{\sum_q\ p(O|q)\ P(q)} \quad (1)$$

First, we extracted phoneme duration from native English (subset of TIMIT (Garofolo, 1993)) data using the alignment module (in Section 2.2.1). We then construct Gaussian distributions specific to each phoneme $\mathbf{p}$ denoted as $\mathbf{D}_{\mathbf{p}}$ to later use it in the IntraVerbalPA framework. Within the framework, using the pre-extracted distribution, we compute the GoPD as follows:

$$GoPD(d_t) = log(P_{D_{p_t}}(d_t)) \quad (2)$$

for a given duration d_t corresponding to a L2-phoneme p_t.

In Figure 2, we present an illustration featuring two phonemes duration distributions, 'V' and 'OY'. Notably, 'OY' exhibits a relatively higher mean duration compared to 'V', which aligns with our expectations since 'V' is a vowel and 'OY' is a consonant. However, it's worth noting that 'V' displays a smaller standard deviation. This characteristic makes 'V' more sensitive to long duration, potentially signaling elongation which will be reflected in the $GoPD(d_t)$.

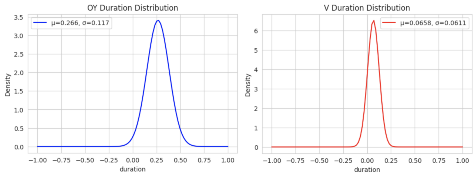

Figure 2: OY vs V duration distribution (*ms*).

Features	Description	Relevance
Loudness	Estimate of perceived signal intensity from an auditory spectrum	Intonation
AlphaRatio	Ratio of the summed energy from 50-1000 Hz and 1-5 kHz. Represents the high-frequency content and the spectral balance.	Intonation
Pitch	logarithmic F0 on a semitone frequency scale	Intonation, Confidence and Expressiveness
JitterLocal	deviations in individual consecutive F0 period lengths	Intonation, Confidence and Expressiveness

Table 1: Selected Frame-level features and their relevance

2.2.3 Frame-level features

In Figure 1, frame-level features F_{nv} are obtained using OpenSmile. It offers 18 low-level descriptors based on the eGeMAPS set (Eyben et al., 2015), including key attributes such as Loudness, AlphaRatio, Pitch, and Jitter Local, along with their corresponding derivatives, as shown in Table 1. Energy is an important feature of speech detection, the energy distribution may be related to the intonation property (Chao et al., 2022), we modeled Energy by Loudness, AlphaRatio, and their derivatives. Pitch provides acoustic cues for a speaker's intonation (Zhang et al., 2021a), confidence, and expressiveness, we present that using logarithmic F0, and JitterLocal.

2.2.4 Utterance-level features

In Figure 1, utterance-level features U_{nv} are also obtained using OpenSmile. In this case, functionals feature levels based on the ComParE set are employed, providing a rich set of up to 6373 features. We explore three strategies for representing U_{nv} using these features:

1. Represent U_{nv} with the complete set of 6373 features, denoted as U_{nv} (#6373).

2. Choose a subset of features and their derivatives detailed in Table 1, resulting in U^s_{nv} (#395).

3. Utilize feature selection using *sklearn.feature_selection.SelectFromModel* through a random forest-trained model to obtain U^m_{nv} (#1590).

2.3 PhoneCue Encoder Module

The PhoneCue Encoder takes as input a sequence $Y = y_1, y_2, \ldots, y_m$ representing parsed canonical phoneme sequence, then to an embedding layer with dimension D : 41. These embeddings are projected using a feedforward operation (with dimension D : 24), resulting in the intermediate feature vector $\tilde{\mathbf{p}}_{\mathbf{nv}}$.

Subsequently, we vertically concatenate this intermediate feature vector **p̃nv** with other relevant components, including GoPD and $\tilde{\mathbf{F}}_{\mathbf{nv}}$ (combined as frame-level features $\tilde{\mathbf{F}}^*_{\mathbf{nv}}$)

Finally, the $\mathbf{r}_{\mathbf{nv}}$ is processed through a Bi-LSTM with dimension (D : 512), resulting in the feature representation $\mathbf{p}_{\mathbf{nv}}$ (D : 1024) capturing the paralinguistic and phonetic cues present in the utterance.

2.4 Projection Block

The $\mathbf{p}_{\mathbf{nv}}$ and the contextualized acoustic representations C_t are then passed to a attention layer that takes $\mathbf{p}_{\mathbf{nv}}$ as query and value, and C_t as key, resulting in the final feature representation $\alpha_{\mathbf{Cnv}}$ (D : 1024)

$$\alpha_{\mathbf{Cnv}} = Attention(K = C_t, Q = \mathbf{p}_{\mathbf{nv}}, V = \mathbf{p}_{\mathbf{nv}}) \tag{3}$$

The embeddings C_t and α_{Cp} (D : 1024) are then concatenated with utterane-level features $\mathbf{U}_{\mathbf{nv}}$, resulting in: $F = C_t + \alpha_{Cpn} + \mathbf{U}_{\mathbf{nv}}$ (4)

The resulting F is then parsed to Bi-LSTM (D : 512), and gets concatenated with the residual utterance-level features $\mathbf{U}_{\mathbf{nv}}$ giving: $\tilde{F} = BiLSTM(F) + \mathbf{U}_{\mathbf{nv}}$ which promotes utterance-level feature reuse. Following, $\tilde{F}$ is then passed to two separate projection layers $Projection_f$, $Projection_p$ of (D : 11), for respective Fluency and Prosodic score classification.

3 Experimental Setup

3.1 Datasets

For the study, we used the widely used Speechocean762 (Zhang et al., 2021b) an extensive dataset specifically designed for pronunciation assessment. The dataset comprises a total of 5,000 English utterances obtained from 250 non-native speakers.

3.2 Model Training and Parameters

The models are optimized using Adam optimizer (Kingma and Ba, 2017) for 25 epochs with early

Exp	$\mathbf{C_t}$	$\mathbf{p_{nv}}$	$\mathbf{F_{nv}}^*$	$\mathbf{U_{nv}}$	**Prosodic**	**Fluency**
			Baselines			
1.I	✓	✗	✗	✗	0.7204	0.7200
1.II	✓	✓	✗	✗	0.7040	0.7092
			Proposed Setups			
1.III	✓	✗	✓	✗	0.7769	0.7740
1.IV	✓	✗	✗	✓	0.7493	0.7452
1.V	✓	✓	✓	✗	0.7429	0.7519
1.VI	✓	✓	✗	✓	0.7372	0.7375
1.VII	✓	✗	✓	✓	0.7689	0.7661
1.VIII	✓	✓	✓	✓	0.7488	0.7481

Table 2: Reported PCC. $\mathbf{F_{nv}}$ converted phoneme-level representation of para-linguistic cues, $*$ including GoPD Goodness of phonemic-duration, $\mathbf{U_{nv}}$: full utterance-level para-linguistic cues representation ($\mathbf{U^f_{nv}}$ #6373). '✓': Feature is included, '✗': Feature is not included

Exp	$\mathbf{C_t}$	$\mathbf{p_{nv}}$	$\mathbf{F_{nv}}^*$	**Prosodic**	**Fluency**
		Experiments with $\mathbf{U^s_{nv}}$ of #395			
2.I	✓	✗	✗	0.7400	0.7407
2.II	✓	✓	✗	0.7507	0.7478
2.III	✓	✗	✓	0.7681	0.7649
2.IV (IntraVerbalPA)	✓	✓	✓	0.7835	0.7851
		Experiments with $\mathbf{U^m_{nv}}$ of #1590			
3.I	✓	✓	✗	0.7327	0.7364
3.II	✓	✓	✗	0.7403	0.7458
3.III	✓	✗	✓	0.7611	0.7617
3.IV	✓	✓	✓	0.7748	0.7709

Table 3: Reported PCC, using $\mathbf{U^s_{nv}}$ and $\mathbf{U^m_{nv}}$ Utterance level features

stopping criterion ($= 3$). The initial learning rate is set to 1×10^{-4}, with a batch size of 32. Following literature, we reported the Pearson Correlation Coefficient (PCC).

4 Results and Discussion

Effectiveness of Proposed Features Table 2 illustrates the effectiveness of the proposed features, namely $\mathbf{p_{nv}}$, $\mathbf{F^*_{nv}}$, and $\mathbf{U_{nv}}$, across various configurations. As depicted in Table 2, the incorporation of these feature combinations consistently results in a notable enhancement in PCC, outperforming the traditional approach of modeling via fine-tuning a pre-trained model to the task (1.I, 1.II), with and without encoded canonical phoneme embedding.

Notably, the optimal outcome is observed when utilizing frame-level features $\mathbf{F_{nv}}$. Significantly, we observed a reduction in PCC by up to 2% upon the inclusion of $\mathbf{p_{nv}}$. Interpreting this decline is challenging in light of existing literature that underscores the effectiveness of incorporating reference phoneme embedding in mispronunciation detection and pronunciation assessment pipelines (Gong et al., 2022; Chao et al., 2022; Fu et al., 2021; Ryu et al.). We think that not including $\mathbf{p_{nv}}$ with frame features in the **PhoneCue module** makes it hard to match up automatically with $\mathbf{C_t}$ features.

Effectiveness of Proposed Selection of Utterance features Table 3 displays the outcomes of various experimental setups utilizing $\mathbf{U^s_{nv}}$ and $\mathbf{U^m_{nv}}$. A consistent trend is observed in both setups, where the PCC incrementally increases with the addition of more proposed features. While the results exhibit marginal differences of less than 0.1% improvement across different utterance features, the selected features $\mathbf{U^s_{nv}}$ yield the optimal outcome, achieving a PCC of 78.35% in Prosody and 78.51% in Fluency outperforming the proposed setup 1.III, 1.VII shown in Table 2 and 3.IV in Table 3. These latter results are deemed the best, affirming the effectiveness of our proposed IntraVerbalPA.

PCC	**Prosodic**	**Fluency**
Contemporary and Proposed Work		
Raw Speech (C_t) (Ryu et al.)	65.00%	65.20%
Wav2vec-large (Kim et al., 2022)	72.00%	72.00%
HiPAMA (Do et al., 2023)	75.10%	74.90%
GOPT (Gong et al., 2022)	76.00%	75.30%
Joint-CAPT-L1 (Ryu et al.)	77.30%	77.50%
Hubert-large-finetuned (Kim et al., 2022)	77.00%	78.00%
MultiPA [Multi-Task PA] (Chen et al., 2023)	78.70%	79.70%
3M (Chao et al., 2022)	82.70%	82.80%
HierarchicalPA (Do et al., 2023)	83.60%	84.30%
IntraVerbalPA (Proposed)	**78.35%**	**78.51%**

Table 4: Reported PCC, for the prior and contemporary works; and our best proposed result IntraVerbalPA

Comparison to Prior Studies: In comparison to contemporary models (see Table 4), the IntraVerbalPA performs comparably with the MultiPA (Chen et al., 2023), and Joint-CAPT (Ryu et al.). While MultiPA and Joint-CAPT operate in a multi-task setup context using either additional features or external L2-Artic.

5 Conclusion

We introduce the IntraVerbalPA framework, enriched with both fine-grained and abstract paralinguistic cues along with the conventional speech and phoneme representation for modeling pronunciation assessment system. Moreover, we propose a new metric to effectively model durtion distribution within the framework. Our reported results validate the importance of individual components of the framework, and demonstrate the efficacy of the IntraVerbalPA.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2022. 3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 575–582.

Yu-Wen Chen, Zhou Yu, and Julia Hirschberg. 2023. Multipa: a multi-task speech pronunciation assessment system for a closed and open response scenario. *Preprint*, arXiv:2308.12490.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *Preprint*, arXiv:2006.13979.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Hierarchical pronunciation assessment with multi-aspect attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Kaiqi Fu, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin. 2021. A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques. *arXiv preprint arXiv:2104.08428*.

John S Garofolo. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266.

Wenping Hu, Yao Qian, and Frank K. Soong. 2015. An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. In *Proc. Speech and Language Technology in Education (SLaTE 2015)*, pages 71–76.

Yassine Kheir, Ahmed Ali, and Shammur Chowdhury. 2023. Automatic pronunciation assessment - a review. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8304–8324, Singapore. Association for Computational Linguistics.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic pronunciation assessment using self-supervised speech representation learning. *Preprint*, arXiv:2204.03863.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Mike Levy and Glenn Stockwell. 2013. *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

Binghuai Lin and Liyuan Wang. 2023. Exploiting information from native data for non-native automatic pronunciation assessment. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 708–714.

Binghuai Lin, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang. 2020. Automatic Scoring at Multi-Granularity for L2 Pronunciation. In *Proc. Interspeech 2020*, pages 3022–3026.

Diane Litman, Helmer Strik, and Gad S Lim. 2018. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309.

Hyungshin Ryu, Sunhee Kim, and Minhwa Chung. A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Mu Yang, Kevin Hirschi, Stephen D Looney, Okim Kang, and John HL Hansen. 2022. Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. *arXiv preprint arXiv:2203.15937*.

Huayun Zhang, Ke Shi, and Nancy F Chen. 2021a. Multilingual speech evaluation: Case studies on english, malay and tamil. *arXiv preprint arXiv:2107.03675*.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021b. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *arXiv preprint arXiv:2104.01378*.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.