Prosody Detection improves Pretrained Automatic Speech Recognition

Anonymous ACL submission

Abstract

We show the performance of Automatic Speech Recognition (ASR) systems that use semi-supervised speech representations can be be boosted by a complimentary prosody detection module, by introducing a joint ASR and prosody detection model. The prosody detection component of our model achieves a significant improvement on the state-of-the-art for the task, closing the gap in F1-score by 41%. Additionally, the ASR performance in joint training decreases WER by 28.3% on LibriSpeech, under limited resource fine-tuning. With these results, we show the importance of extending pretrained speech models to retain or relearn important prosodic cues.

1 Introduction

Models based on self-supervised speech representations have in recent years claimed state-of-the-art performance in ASR (Baevski et al., 2020; Hsu et al., 2021). Moreover, they have permitted to bypass both a heavy speech-science informed featurisation component, as well a language dependent acoustic dictionary resource writing component. In doing so, such models have become seemingly less human reliant during development, provided adequate quantities of raw speech data and computational resources.

However, the training techniques for these self-supervised speech models do not reveal what exactly is deemed important by these models and later retained within their output speech representations. Subsequent studies have since introduced benchmarks and metrics to analyse the linguistic knowledge of these models at different levels, mostly from the point of view of assessing the existence of this linguistic knowledge. For example, the Zero-Resource Speech challenges[1] provide tests beds to analyse phonetic, lexical, syntactic, and semantic level book-keeping of these representations (Dunbar et al., 2017, 2020; Nguyen et al., 2021). More recently, ProsAudit was introduced to provide a similar book-keeping of the prosodic information retained in these speech representations (de Seyssel et al., 2023). However, these studies and corresponding benchmarks do not elucidate whether refocusing these SSL representations to retain more of the original linguistic signal could correspond to better performance downstream in basic but central speech tasks like ASR.

One main bottleneck to carrying out such a study is the sheer scarcity of datasets with prosodic annotations, and the previous expense in generating these annotations using trained linguists. One can, however, envisage a scheme for obtaining simple prosodic annotations for "important" words in an utterance from non-specialists, and even for non-written languages, by which annotators simply press a button when important segments of an utterance are heard. Still, currently, the only existing datasets for English (and any other language) are relatively small–the largest being the Boston University Radio News Corpus for English with 11 hours of data (Ostendorf et al., 1995). Is there any benefit for spoken language understanding tasks in extending current and/or developing new prosody datasets? To assess this question, one must investigate the role of prosody in these tasks.

Contributions. In this paper we study the role of prosodic information, specifically focusing on pitch accents, in ASR. Our contributions are as follows.

1. We streamline and significantly boost the performance of the current state-of-the-art model for pitch accent detection.
2. We present a multi-task model for integrating pitch accent detection into the ASR task, which improves the performance for ASR in limited resource settings.

[1] https://zerospeech.com

3. We then automatically annotate the pitch accents of a small dataset using self-training, and then apply it in our proposed joint model, achieving even further ASR performance boosts.

2 Related work

Prosody Detection. There is a long line of research on automatic prosody detection (for example, (Taylor, 1995; Rosenberg et al., 2015; Shahin et al., 2016; Li et al., 2018; Stehwien et al., 2020; Sabu et al., 2021)). With the advent of pretrained speech models, and in particular, wav2vec (Schneider et al., 2019) and wav2vec2 (Baevski et al., 2020), a new line of systems that builds on self-supervised speech representations has achieved the state-of-the-art in detecting prosodic boundaries in Czech broadcast news recordings (Kunešová and Řezáčková, 2022) and in pitch events and intonation phrase boundaries in English broadcast news (Zhai and Hasegawa-Johnson, 2023). This latter model, called wav2TOBI, forms the point of departure for our multitask system presented here. The question left open by these models, and others is whether these self-supervised representations adequately account for prosody, which has been shown to aid in an array of linguistic tasks like SLU (Nöth et al., 2002; Shriberg and Stolcke, 2004; Shriberg et al., 1998; Rajaa, 2023; Wei et al., 2022), and parsing (Tran et al., 2017; Gregory et al., 2004; Kahn et al., 2005; Dreyer and Shafran, 2007; Kahn and Ostendorf, 2012; Price et al., 1991; Beckman, 1996). Or whether a specialised module should intervene and boost the prosodic signal for better performance.

Prosody with ASR. In this paper, we are particularly interested in whether refocusing speech pretrained models on prosody might aid in performance for ASR. Prosody has been previously shown to be of importance to ASR, both as engineered features, as well as through learning from prosody annotated datasets (Silverman et al., 1992; Ostendorf et al., 2003; Hirose and Minematsu, 2004; Hirschberg et al., 2004; Hasegawa-Johnson et al., 2005; Ananthakrishnan and Narayanan, 2007; Vicsi and Szaszák, 2010; Chen et al., 2012; Kathania et al., 2020; Hasija et al., 2022; Coto-Solano, 2021). However, to our knowledge, there is no research that builds on pretrained speech models, whose application to prosody detection and ASR has resulted in the state-of-the-art performance.

State-of-the-art ASR A summary of state-of-the-art performance in ASR over the Librispeech dataset is given in the appendix (Table 3). For this paper, for model comparability, we focus on the wav2vec2 model, which is the pretrained model fine-tuned by the wav2TOBI model for prosody detection, and which is the model on which we base our system presented here.

3 Modelling prosody and ASR

3.1 Datasets

Our research uses the BURNC (Ostendorf et al., 1995), Librispeech (Panayotov et al., 2015) and Libri-light (Kahn et al., 2020) corpora.

The BURNC dataset is a broadcast news-style read speech corpus which contains 11 hours of speech, sourced from 7 different speakers (3 female and 4 male), It consists of audio snippets with their transcriptions, phonetic alignments, parts-of-speech tags and prosodic labels. We used 75% of the data in this dataset for training, 15% for development and 10% for testing. Because multiple readers may have read the same news story in BURNC, we ensure that no news stories appearing in the test set also occur in the training set.

We represent pitch accent labels from the BURNC following the binary labelling strategy presented in (Zhai and Hasegawa-Johnson, 2023). We assign positive labels to time-frames corresponding to audio segments labelled in the BURNC as having pitch accents, and negative labels otherwise.

Following (Zhai and Hasegawa-Johnson, 2023), we preprocess the BURNC audios by splitting them into overlapping clips of 20s, at 10s intervals.

The Librispeech dataset consists of 1000 hours of audio samples sourced from the LibriVox Project. In our work, the dev-clean and test-clean data subsets were used for model development and evaluation, respectively.

The Libri-light dataset is made up of 60,000 hours of audio and, similarly to the Librispeech corpus, was also sourced from the LibriVox Project. We used the Libri-light limited resource training data subsets, namely, train-1h (LS1), which consists of 1 hour of labelled audio data.

3.2 A joint model for prosody and ASR

Our proposed system uses prosody annotations to jointly learn pitch accent detection and automatic speech recognition (cf Figure 1).

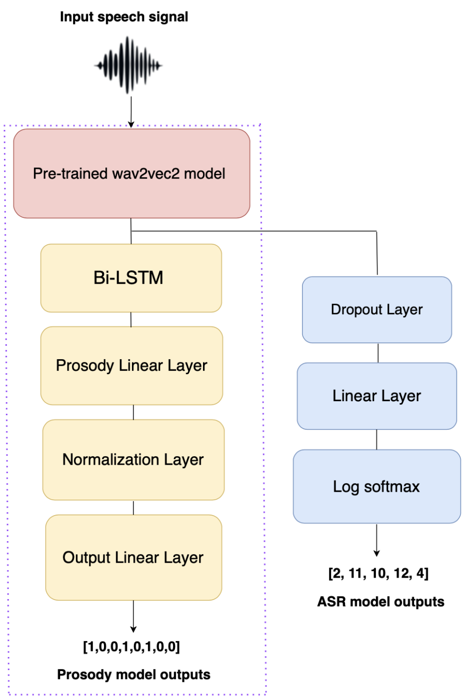

Figure 1: Joint Prosody-ASR Model

Model	Tol	Prec	Rec	F1
wav2TOBI	0 ms	0.13	0.11	0.12
	40 ms	0.70	0.61	0.65
	80 ms	0.87	0.74	0.79
	100 ms	0.89	0.76	0.81
Ours Prosody	0 ms	0.37	0.36	0.36
	40 ms	0.82	0.8	0.81
	80 ms	0.89	0.86	0.87
	100 ms	0.9	0.87	0.88
Ours Semi-Sup	0 ms	0.49	0.48	0.48
	40 ms	0.85	0.82	0.83
	80 ms	0.92	0.88	0.9
	100 ms	0.93	0.89	0.9

Table 1: Prosody detection system performance at varying levels of error tolerance (Tol) in milliseconds. Our basic model (Ours Prosody) outperforms the state of the art wav2TOBI system in pitch accent detection. Our semi-supervised training approach (Ours Semi-Sup) further improves performance.

For both ASR and prosody detection, raw audio input is sent through the pretrained wav2vec2 model (Baevski et al., 2020) with a language modelling head on top for Connectionist Temporal Classification for the ASR task. For pitch accent detection, we built upon the prosodic event detection model proposed by Zhai and Hasegawa-Johnson (2023), wav2TOBI. In wav2TOBI, wav2vec2 timestep representations are concatenated with fundamental frequency features and fed through a BiLSTM, followed by a classification layer, with mean-squared loss. Our model streamlines wav2TOBI in the sense that we no longer require fundamental frequency features and make pure use of wav2vec2 output representations. On the other hand, we introduce an extra linear layer followed by layer normalisation before the classification layer.

We train our proposed joint model by minimising a joint loss function $\mathcal{L}_j = \mathcal{L}_{asr} + \mathcal{L}_{pad}$, which is the combination of the ASR model loss $\mathcal{L}_{asr}$ and pitch accent detection model loss $\mathcal{L}_{pad}$.

Results for prosody detection. In the single task setting, for prosody detection, these simple changes result in significant improvements in pitch accent detection performance over wav2TOBI, even without recourse to the additional fundamental frequency features (Table 1).

3.3 Semi-supervised Prosodic Event Detection

Our joint prosody-ASR modelling is limited to datasets where these prosodic labels are available. In order to address ASR performance for a dataset like LibriSpeech, where prosodic labels are unavailable, we resort to semi-supervision–specifically, self-training with model voting. In these experiments, we focus on the smaller Libri-light train-1h (LS1) dataset in order to minimise the possible extrapolation error of a larger dataset.

We partitioned the BURNC train set into three subsets as possible hold-outs. For each hold-out subset, we used the remaining 2/3s of the original train set to retrain a new model. We used each of the three models to obtain three predicted labels for each instance of the LS1 train set, and retained the majority class label of these for each instance. The prosody labelled version of LS1 was then added to the full BURNC train set, and then checked over the BURNC test set for performance gains. If there were gains, we repeated the process now with the prosody labelled LS1 as part of the partitioning step, replacing the labels of LS1 at each iteration. Otherwise the process halts. Our training process halted after 4 iterations.[2]

Results for prosody detection. The single task results for this approach on prosody detection are

[2]Note that we tried a number of different self-training techniques, but this simple voting technique worked the best.

given in Table 1. We observe that across all measures and error tolerances, this method improves performance and achieves, to our knowledge, the current state-of-the-art.

4 Experimental setup and results

We use the base-960h wav2vec2 pretrained model[3]. Our models all are trained for 30,000 steps, using default parameters.[4] Results for word and character error rates (respectively WER and CER) are given in Table 2. All models were fine-tuned for ASR (resp. ASR and prosody jointly) on LS1, and thereafter possibly fine-tuned on BURNC (indicated by ft BURNC). In the fine-tuning process, following Baevski et al. (2020), the pretrained model remains frozen during the first 15K steps, after which the entire model is trained for the remaining 15K steps. The feature encoder remains frozen throughout fine-tuning. For the Prosody-ASR model, we use the prosody labelled version of LS1 outlined above.

	LibriSpeech		BURNC	
Model	**WER**	**CER**	**WER**	**CER**
ASR-only	6.0	**1.0**	23.0	7.0
ASR-only (ft BURNC)	4.9	**1.0**	**12.0**	**4.0**
Prosody-ASR	**4.3**	1.1	20.0	7.7
Prosody-ASR (ft BURNC)	**4.3**	1.1	20.0	7.0

Table 2: WER and CER results for ASR.

Results. We observe that while there is no great change to CER scores, the joint Prosody-ASR model improves both WER on LibriSpeech and BURNC test data by 28.3% and 13% respectively, showing that a refocus of the wav2vec2 representations on prosody helps to improve ASR performance over LibriSpeech. Interestingly, our semi-supervised approach yields worst WER for BURNC than bypassing prosodic labels for fine-tuning on both LS1 and BURNC. We posit that this may be due to the noisiness of the inferred prosodic labels in the LS1 dataset specifically for BURNC. We therefore tried fine-tuning the joint model solely on BURNC; however both WER and CER increased to 29.0 and 9.0 respectively. This may be due to the small size of the BURNC dataset in combination with the quantity of information that must be learned in the joint model.

[3] https://huggingface.co/facebook/wav2vec2-base-960h

[4] Default parameters are from https://huggingface.co/docs/transformers/en/model_doc/wav2vec2#transformers.Wav2Vec2ForCTC.

5 Error analysis and discussion

We have shown above that pitch accent detection is useful for improving the performance of pretrained speech models in ASR tasks within limited resource scenarios. However, even though we improve upon the WER in most of the experiments that we perform with our proposed joint model, we notice that experiments that involve the BURNC dataset tend to on average have higher CER scores. We list two reasons for this phenomenon below and discuss their impact.

Pre-processing mismatches and audio truncation. During the pre-processing of the transcriptions for the BURNC dataset, we transform the text into their uppercase representations and remove all punctuation marks that are not consequential in determining word meaning. For instance, given a word "CHIEF'S", we do not remove the apostrophe (') during pre-processing to form the word "CHIEFS" since doing so changes the inherent meaning of the word. Another example is "S.J.C's", we do not represent it as "SJCS". Even though this text pre-processing approach is well-warranted, it leads to higher CER scores during ASR since the BURNC dataset is filled with a plethora of acronyms, hyphenated and contracted words.

Following the data pre-processing approach utilised in (Zhai and Hasegawa-Johnson, 2023), we split our audios into overlapping clips of 20s, at intervals of 10s, for input to the wav2vec2 model. This however leads to the truncation of words in initial or final position. As a result of this, some of the words that are predicted by the model are incomplete and this leads to a higher CER score.

6 Conclusion

In this paper we have presented an approach for leveraging prosodic information to improve the performance of a pretrained speech model in a limited resource scenario. The results from our experiments demonstrate that re-focusing self-supervised speech models on supra-segmental speech cues such as prosody could lead to significant performance gains in downstream tasks.

7 Limitations

All experiments were carried out under the limited resource setting, with little fine-tuning data, due to the requirement of our method to use prosodic labels. More work is required to investigate the real impact when fine-tuning with larger ASR datasets.

Also, for prosodic cues, we only used pitch-accent, and with hard labels (0 or 1). It is not clear whether other aspects of prosody would also be important for ASR. This question remains open.

References

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2007. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–873. IEEE.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pretraining for speech recognition. *arXiv preprint arXiv:1911.03912*.

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.

Mary E Beckman. 1996. The parsing of prosody. *Language and cognitive processes*, 11(1-2):17–68.

Sin-Horng Chen, Jyh-Her Yang, Chen-Yu Chiang, Ming-Chieh Liu, and Yih-Ru Wang. 2012. A new prosody-assisted mandarin asr system. *Ieee transactions on audio, speech, and language processing*, 20(6):1669–1684.

Rolando Coto-Solano. 2021. Explicit tone transcription improves asr performance in extremely low-resource languages: A case study in bribri. In *Proceedings of the first workshop on natural language processing for Indigenous languages of the Americas*, pages 173–184.

Maureen de Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. 2023. ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proceedings of INTERSPEECH 2023*, pages 2963–2967.

Markus Dreyer and Izhak Shafran. 2007. Exploiting prosody for pcfgs with latent annotations. In *INTERSPEECH*, pages 450–453. Citeseer.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The zero resource speech challenge 2020: Discovering discrete subword and word units. In *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Michelle Gregory, Mark Johnson, and Eugene Charniak. 2004. Sentence-internal prosody does not help parsing the way punctuation does. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 81–88.

Mark Hasegawa-Johnson, Ken Chen, Jennifer Cole, Sarah Borys, Sung-Suk Kim, Aaron Cohen, Tong Zhang, Jeung-Yoon Choi, Heejin Kim, Taejin Yoon, et al. 2005. Simultaneous recognition of words and prosody in the boston university radio speech corpus. *Speech Communication*, 46(3-4):418–439.

Taniya Hasija, Virender Kadyan, Kalpna Guleria, Abdullah Alharbi, Hashem Alyami, and Nitin Goyal. 2022. Prosodic feature-based discriminatively trained low resource speech recognition system. *Sustainability*, 14(2):614.

Keikichi Hirose and Nobuaki Minematsu. 2004. Use of prosodic features for speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech communication*, 43(1-2):155–175.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Jeremy G Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech.

In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 233–240.

Jeremy G Kahn and Mari Ostendorf. 2012. Joint reranking of parsing and word recognition with automatic segmentation. *Computer Speech & Language*, 26(1):1–19.

Hemant Kathania, Mittul Singh, Tamás Grósz, and Mikko Kurimo. 2020. Data augmentation using prosody and false starts to recognize non-native children's speech. *arXiv preprint arXiv:2008.12914*.

Marie Kunešová and Markéta Řezáčková. 2022. Detection of prosodic boundaries in speech using wav2vec 2.0. In *International Conference on Text, Speech, and Dialogue*, pages 377–388. Springer.

Kun Li, Shaoguang Mao, Xu Li, Zhiyong Wu, and Helen Meng. 2018. Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks. *Speech Communication*, 96:28–36.

Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert. 2020. slimipl: Language-model-free iterative pseudo-labeling. *arXiv preprint arXiv:2010.11524*.

Shaoshi Ling and Yuzong Liu. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2021. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Proceedings of Interspeech 2021*, pages 1574–1578.

Elmar Nöth, Anton Batliner, Volker Warnke, Jürgen Haas, Manuela Boros, Jan Buckow, Richard Huber, Florian Gallwitz, Matthias Nutt, and Heinrich Niemann. 2002. On the use of prosody in automatic dialogue understanding. *Speech Communication*, 36(1-2):45–62.

Mari Ostendorf, Patti J Price, and Stefanie Shattuck-Hufnagel. 1995. The boston university radio news corpus. *Linguistic Data Consortium*, pages 1–19.

Mari Ostendorf, Izhak ShafranÞ, and Rebecca Bates. 2003. Prosody models for conversational speech recognition.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. 2020. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629*.

Patti J Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, 90(6):2956–2970.

Shangeth Rajaa. 2023. Improving end-to-end slu performance with prosodic attention and distillation. *arXiv preprint arXiv:2305.08067*.

Andrew Rosenberg, Raul Fernandez, and Bhuvana Ramabhadran. 2015. Modeling phrasing and prominence using deep recurrent learning. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kamini Sabu, Mithilesh Vaidya, and Preeti Rao. 2021. Cnn encoding of acoustic parameters for prominence detection. *arXiv preprint arXiv:2104.05488*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 3465–3469. ISCA.

Mostafa Ali Shahin, Julien Epps, and Beena Ahmed. 2016. Automatic classification of lexical stress in english and arabic languages using deep learning. In *Interspeech*, pages 175–179.

Elizabeth Shriberg and Andreas Stolcke. 2004. Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing*, pages 105–114. Springer.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

Kim EA Silverman, Eleonora Blaauw, Judith Spitz, and John F Pitrelli. 1992. Towards using prosody in speech recognition/understanding systems: Differences between read and spontaneous speech. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Sabrina Stehwien, Antje Schweitzer, and Ngoc Thang Vu. 2020. Acoustic and temporal representations in convolutional neural network models of prosodic events. *Speech Communication*, 125:128–141.

Paul A Taylor. 1995. Using neural networks to locate pitch accents.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. *arXiv preprint arXiv:1704.07287*.

Klára Vicsi and György Szaszák. 2010. Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5):413–426.

Kai Wei, Dillon Knox, Martin Radfar, Thanh Tran, Markus Müller, Grant P Strimel, Nathan Susanj, Athanasios Mouchtaris, and Maurizio Omologo. 2022. A neural prosody encoder for end-to-end dialogue act classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7047–7051. IEEE.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020. Iterative pseudo-labeling for speech recognition. *arXiv preprint arXiv:2005.09267*.

Wanyue Zhai and Mark Hasegawa-Johnson. 2023. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2748–2752. Publisher Copyright: © 2023 International Speech Communication Association. All rights reserved.; 24th International Speech Communication Association, Interspeech 2023 ; Conference date: 20-08-2023 Through 24-08-2023. [link].

8 Appendix

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
10-min labeled						
DiscreteBERT (Baevski et al., 2019)	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE (Baevski et al., 2020)	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE (Baevski et al., 2020)	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE (Baevski et al., 2020)	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE (Hsu et al., 2021)	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE (Hsu et al., 2021)	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE (Hsu et al., 2021)	LL-60k	Transformer	4.3	7.0	4.7	7.6
HUBERT X-LARGE (Hsu et al., 2021)	LL-60k	Transformer	4.4	6.1	4.6	6.8
1-hour labeled						
DeCoAR 2.0 (Ling and Liu, 2020)	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT (Baevski et al., 2019)	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE (Baevski et al., 2020)	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE (Baevski et al., 2020)	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE (Hsu et al., 2021)	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE (Hsu et al., 2021)	LL-60k	Transformer	2.6	4.9	2.9	5.4
HUBERT X-LARGE (Hsu et al., 2021)	LL-60k	Transformer	2.6	4.2	2.8	4.8
10-hour labeled						
SlimIPL (Likhomanenko et al., 2020)	LS-960	4-gram + Transformer	5.3	7.9	5.5	9.0
DeCoAR 2.0 (Ling and Liu, 2020)	LS-960	4-gram	-	-	5.4	13.3
DiscreteBERT (Baevski et al., 2019)	LS-960	4-gram	5.3	13.2	5.9	14.1
wav2vec 2.0 BASE (Baevski et al., 2020)	LS-960	4-gram	3.8	9.1	4.3	9.5
wav2vec 2.0 LARGE (Baevski et al., 2020)	LL-60k	Transformer	2.4	4.8	2.6	4.9
HUBERT BASE (Hsu et al., 2021)	LS-960	4-gram	3.9	9.0	4.3	9.4
HUBERT LARGE (Hsu et al., 2021)	LL-60k	Transformer	2.2	4.3	2.4	4.6
HUBERT X-LARGE (Hsu et al., 2021)	LL-60k	Transformer	2.1	3.6	2.3	4.0
100-hour labeled						
IPL (Xu et al., 2020)	LL-60k	4-gram + Transformer	3.19	6.14	3.72	7.11
SlimIPL (Likhomanenko et al., 2020)	LL-60k	4-gram + Transformer	2.2	4.6	2.7	5.2
Noisy Student (Park et al., 2020)	LS-860	LSTM	3.9	8.8	4.2	8.6
DeCoAR 2.0 (Ling and Liu, 2020)	LS-960	4-gram	-	-	5.0	12.1
DiscreteBERT (Baevski et al., 2019)	LS-960	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE (Baevski et al., 2020)	LS-960	4-gram	2.7	7.9	3.4	8.0
wav2vec 2.0 LARGE (Baevski et al., 2020)	LL-60k	Transformer	1.9	4.0	2.0	4.0
HUBERT BASE (Hsu et al., 2021)	LS-960	4-gram	2.7	7.8	3.4	8.1
SlimIPL (Likhomanenko et al., 2020)	LL-60k	Transformer	1.8	3.7	2.1	3.9
HUBERT X-LARGE (Hsu et al., 2021)	LL-60k	Transformer	1.7	3.0	1.9	3.5

Table 3: Comparison of ASR model performance on the Librispeech dataset (Hsu et al., 2021)