ON THE UTILITY OF SELF-SUPERVISED MODELS FOR PROSODY-RELATED TASKS

Guan-Ting Lin[1†], *Chi-Luen Feng*[1†], *Wei-Ping Huang*[1‡], *Yuan Tseng*[1‡], *Tzu-Han Lin*[1‡], *Chen-An Li*[1‡], *Hung-yi Lee*[1], *Nigel G. Ward*[2]

[1]National Taiwan University, Taiwan
[2]University of Texas at El Paso, USA

ABSTRACT

Self-Supervised Learning (SSL) from speech data has produced models that have achieved remarkable performance in many tasks, and that are known to implicitly represent many aspects of information latently present in speech signals. However, relatively little is known about the suitability of such models for prosody-related tasks or the extent to which they encode prosodic information. We present a new evaluation framework, "SUPERB-prosody," consisting of three prosody-related downstream tasks and two pseudo tasks. We find that 13 of the 15 SSL models outperformed the baseline on all the prosody-related tasks. We also show good performance on two pseudo tasks: prosody reconstruction and future prosody prediction. We further analyze the layerwise contributions of the SSL models. Overall we conclude that SSL speech models are highly effective for prosody-related tasks. We release our code[1] for the community to support further investigation of SSL models' utility for prosody.

Index Terms— Speech Self-Supervised Learning, Representation Learning, Pretrained Models, Prosody, Pragmatics

1. INTRODUCTION

Self-supervised Learning (SSL) has revolutionized research in many areas of artificial intelligence, including speech processing. SSL pre-trained speech models have shown remarkable performance and generalizability across a wide range of tasks [1, 2, 3]. However, we do not currently have a good understanding of what knowledge these models capture nor of the limits of their power. This is true in particular for the prosodic aspects of speech.

The speech signal contains not only lexical but prosodic information. Broadly speaking, the latter has three realms of function: paralinguistic, phonological, and pragmatic. Paralinguistic functions, such as marking speaker identity and expressing emotion, are largely conveyed by prosodic settings that are stable over the span of many utterances, and are often evident from any sample of just a few syllables. Phonological functions, notably marking the identity of syllables and words with tones and stress patterns, are largely conveyed by prosodic features whose temporal occurrence is tightly linked to the units they mark. The utilities of SSL models for these two realms have been demonstrated, by many recent studies using tasks from the SUPERB [1] benchmark, among others.

However, for the third realm, the realm of pragmatic function, the question of the utility of SSL models has remained open. Functions in this realm, include managing turn-taking, marking topic structure and information structure, and expressing engagement, stance, attitude, and intent. These pragmatic functions are especially important in dialog, and we expect that future dialog systems will need more prosodic competence, in order to enable more satisfying user experiences and to support interaction in novel genres and situations [4]. In many cases, these functions are expressed using multistream temporal configurations of low-level prosodic features, where these configurations can last from a few hundred milliseconds to several seconds [5, 6], and may be only loosely aligned with the lexical content. As these configurations are fundamentally different from the forms of prosody in the other two realms, it is an open question whether SSL models are also useful for this realm.

Accordingly our research question is whether pre-trained models have utility for prosody-conveyed pragmatic functions. We investigate this in four ways. First, we assemble a set of prosody-intensive tasks and measure how well pre-trained models support them. Second, we use pitch and energy reconstruction pseudo-tasks to measure how well these models represent prosodic information. Third, we evaluate the utility of these models for the prediction of future pitch and energy. Fourth, we probe the pre-trained models to see in which layers prosodic information is likely represented.

The main contributions of this paper are: 1) The finding that pre-trained SSL models indeed can provide value for prosody-intensive tasks, often reaching state-of-the-art performance. 2) Results for 15 recent SSL models span different model architectures and pre-training objectives. 3) Analysis of the representation of prosody in SSL models, including layerwise analysis. 4) An open-source evaluation framework, SUPERB-prosody, examines the prosodic prowess of SSL models.

[†‡]Equal contribution
[1]https://github.com/JSALT-2022-SSL/superb-prosody

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

2. RELATED WORK

The most directly relevant study [7] aimed primarily to evaluate a model for producing de-identified representations of speech, but includes three aspects that are very relevant to our research question. First, for evaluation purposes, several "spoken language understanding" tasks were selected, of which three of these were both pragmatics-relevant and prosody-intensive. Second, six pre-trained models were tested against this task set, showing various levels of performance. Third, probing their own model, VQP, provided evidence that it was encoding, to some extent, several prosodic features. Taken together these results suggest that the answer to our research question is yes, but the case is not settled for two reasons. First, neither their performance results nor their probing results were compared against non-pre-trained baselines, leaving open the question of whether the pre-trained models were in fact providing any benefit. Second, many aspects of their methods are unclear, and no code is available to enable replication. Thus, we need further investigation, and an open and transparent evaluation framework for SSL models. Very recent work has found that SSL models are helpful for predicting some perceptions of speaking style, but require additional downstream sequence-modeling layers for best performance [8].

Evaluation benchmarks have been critical in supporting and evaluating the rise of SSL in the speech field. NOSS [9] is a benchmark for non-semantic downstream tasks. SUPERB [1] broadly examines SSL models for content, speaker, semantic and paralinguistic aspects, demonstrating that SSL models generalize across diverse downstream tasks. SUPERB-sg [2] enhances the SUPERB benchmark with more challenging semantic and generative tasks. SLUE [10], another recent benchmark, targets spoken language understanding tasks. However, up to now, the speech community lacks an evaluation benchmark/framework to measure the prosodic utility of SSL models.

Analysis of speech SSL models has recently received significant attention, with previous works examining various aspects [11, 12, 13, 7]. [11, 13] mainly focus on analyzing lexical information in SSL models. [12] investigate acoustic, syntactic, and semantic characteristics, but they only experiment on two SSL models and probe only by utterance-wise regression tasks. Accordingly, more focused analysis is needed, especially regarding prosodic information.

3. THE SUPERB-PROSODY FRAMEWORK

This section introduces our tasks — classification, prosody reconstruction, and future value prediction task — and our evaluation framework, including the upstream/downstream setup.

3.1. Classification tasks

To evaluate the pragmatic and prosody-related abilities of pre-trained models, we need a set of "prosody-intensive" tasks,

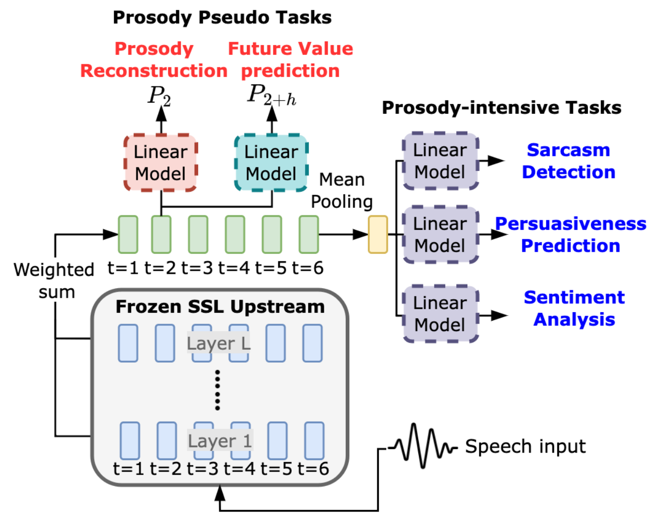

Fig. 1. Diagram of `SUPERB-Prosody` framework. We extract the hidden representations from a frozen SSL model, and lightweight linear models are used for each downstream task. P_i means the value of the rule-based prosodic feature at time frame i.

that is, tasks where it is known that prosodic abilities are useful. We chose three well-curated, open-source, utterance-classification datasets, involving sentiment, sarcasm, and persuasiveness. To briefly describe each task:

Sentiment Analysis (SA) involves detecting the degree of positive or negative feeling in an utterance. While sentiment and emotion are often conflated, and similar methods may work for both tasks, sentiment is less visceral and of greater practical importance. Specifically we chose the CMU-MOSEI [26] corpus, in which each utterance is labeled from -3 to 3, representing the degree of sentiment. Following previous works, we experiment with two settings: binary classification, with the dataset split by the labels in $[-3, 0)$ and $(0, 3]$, and seven-category classification. The evaluation metric is accuracy.

Sarcasm Detection (SarD) is perhaps the most obviously prosody-intensive task, as a mismatch between the lexical content and the prosodic message is frequently the major marker of sarcasm. We chose the MUStARD [27] corpus, in which each utterance has a label of 0 or 1, for sarcasm or non-sarcasm. We follow the speaker-dependent setup in the original paper [27], using five-fold cross-validation for evaluation. The evaluation metric is the F1 score.

Persuasiveness Prediction (PP) is the task of detecting whether a presentation is likely to be convincing to others. Correlates include how pleasant the speaker's voice is and their perceived confidence and dominance, all of which involve prosody. We chose the POM [28] corpus. The labels are 0 or 1, for persuasive or non-persuasive. The evaluation metric is accuracy.

Model	Network	#Params	Stride	Input	Corpus	Pretraining
FBANK	-	0	10ms	waveform	-	-
APC [14]	3-GRU	4.11M	10ms	FBANK	LS 360 hr	F-G
VQ-APC [15]	3-GRU	4.63M	10ms	FBANK	LS 360 hr	F-G + VQ
NPC [16]	4-Conv, 4-Masked Conv	19.38M	10ms	FBANK	LS 360 hr	M-G + VQ
Mockingjay [17]	12-Trans	85.12M	10ms	FBANK	LS 360 hr	time M-G
TERA [18]	3-Trans	21.33M	10ms	FBANK	LS 960 hr	time/freq M-G
modified CPC [19]	5-Conv, 1-LSTM	1.84M	10ms	waveform	LL 60k hr	F-C
wav2vec [20]	19-Conv	32.54M	10ms	waveform	LS 960 hr	F-C
vq-wav2vec [21]	20-Conv	34.15M	10ms	waveform	LS 960 hr	F-C + VQ
DistilHuBERT [22]	7-Conv 2-Trans	23.49M	20ms	waveform	LS 960 hr	KD
wav2vec 2.0 Base [23]	7-Conv 12-Trans	95.04M	20ms	waveform	LS 960 hr	M-C + VQ
wav2vec 2.0 Large [23]	7-Conv 24-Trans	317.38M	20ms	waveform	LL 60k hr	M-C + VQ
HuBERT Base [24]	7-Conv 12-Trans	94.68M	20ms	waveform	LS 960 hr	M-P + VQ
HuBERT Large [24]	7-Conv 24-Trans	316.61M	20ms	waveform	LL 60k hr	M-P + VQ
WavLM Base [25]	7-Conv 12-Trans	94.68M	20ms	waveform	LL 60k hr	M-P + VQ
WavLM Large [25]	7-Conv 24-Trans	316.62M	20ms	waveform	Mix 94k hr	M-P + VQ

Table 1. SSL models examined. #Params includes parameters for both pre-training and inference. LS = LibriSpeech and LS = LibriLight. For the pre-training methods, VQ = vector quantization, F = future, M = masked, G = generation, C = contrastive discrimination, P = token prediction/classification, and KD = knowledge distillation.

3.2. Prosody reconstruction

Prosody Reconstruction (ProP) is a pseudo-task designed to test whether SSL models embed specific prosody features in their hidden representation. As our target, we chose the two most commonly used prosody features, pitch and energy. Given the SSL features, we use a lightweight linear downstream model to predict each. The pitch is represented in log scale, using as targets the values computed by pYAAPT[2]. For energy, we use the librosa toolkit[3], again using a log scale. As data, we use LibriTTS [29], a multi-speaker text-to-speech dataset, and for both features, the evaluation metric is the Mean Square Error (MSE) of the differences. No loss is computed for unvoiced frames, that is, frames where pYAAPT detects no pitch.

3.3. Future value prediction

Future Value Prediction (FVP) is designed to test whether the output from SSL models can predict future prosody. The task setting is similar to the prosody reconstruction task, using pitch and energy features as the prediction targets and using LibriTTS dataset with MSE objective. The task is, given the information from $t = 1$ to the current frame i, to predict the value at $i + h$, where h is the prediction horizon. Four prediction horizons h, namely 0.12, 0.24, 0.50, and 1.00 seconds, are used. Because most SSL speech models are not causal (due to the inclusion of either self-attention or bi-directional connections), we only test causal SSL models or attention-based SSL models for which we can apply an attention mask to avoid cheating with future information[4]. The evaluation metrics are MSE.

[2]http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html

[3]https://github.com/librosa/librosa

3.4. Evaluation framework

As suggested by Figure 1, our framework consists of (1) an upstream SSL speech model and (2) a linear downstream model for probing. Following the procedure of `SUPERB` [1], the parameters of the upstream models are fixed for all the downstream tasks. For each frame of the input, we extract the representations $\mathbf{x_i}$ from each hidden layer i of the upstream model, and we aggregate those hidden representations per utterance into $\mathbf{y} = \sum_i^L w_i \mathbf{x_i}$ where the w_i are trained per task. The resulting $\mathbf{y}$, a two-dimensional matrix (time $\times$ aggregated features), is the input for the downstream model.

3.4.1. Upstream models: Speech SSL models

The SSL models tested are summarized in Table 1. These are a diverse collection, including modified CPC [19], APC [14], VQ-APC [15], NPC [16], TERA [18], vq-wav2vec [21], DistilHuBERT [22], HuBERT [24], wav2vec [20], wav2vec 2.0 [23], and WavLM [25]. The selected SSL models span different network architectures and pre-training objectives.

3.4.2. Downstream models: linear probing model

For the classification tasks, SA, SarD, and PP, the representation $\mathbf{y}$ is mean-pooled along the time axis, forming a dense

[4]We exclude WavLM for this task because it uses a gated relative position bias, so simply modifying the attention mask would cause a model mismatch between training and inference.

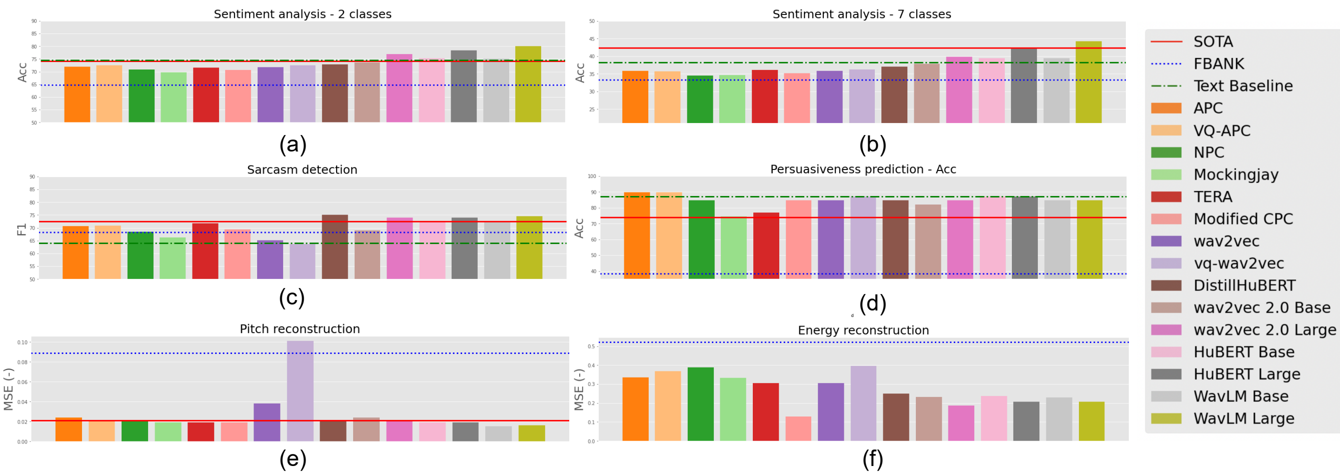

Fig. 2. Results for Sentiment Analysis (SA), Sarcasm Detection (SarD), Persuasiveness Prediction (PP), and Prosody Reconstruction (PR). State-of-the-art (SOTA) values are from [26] for SA, [30] for SarD, [31] for PP, and REAPER[32] for pitch reconstruction. Text-only baselines are only for the prosody-intensive downstream tasks, and thus not available for PR.

vector of dimension (time × aggregated features) to (aggregated features). This vector is then fed to a simple linear model to project from model dimension to 1. The training objective is Cross-Entropy Minimization.

In ProR and FVP, the goal is to predict the fine-grained prosodic information. We use the *frame-level representation* from each time step of $\mathbf{y}$ as the input, and the downstream model is a linear model, projecting from model dimension to 1. MSE minimization is the training objective. We try multiple learning rates for each task ($[1e^{-2}$, $1e^{-3}$, $1e^{-4}$, $1e^{-5}$, $1e^{-6}]$), and report the best performance. The training step is 3000 for SarD and 50000 for the other tasks.

3.4.3. *Baselines*

As a baseline feature set, we use"FBANK," the 80-dimensional Log Mel Filterbank features with delta and delta-delta features (240 dimensions in total), chosen because this is known to work well for many speech tasks. For the classification tasks, FBANK features are average-pooled across each utterance. For pitch reconstruction, we also compare the performance of another high-quality off-the-shelf pitch extractor, Talkin's REAPER. For future value prediction (FVP) we design a baseline, FBANK + RNN, which feeds filterbank features from $t = 1$ to i into a one-layer uni-directional Recurrent Neural Network (RNN) with 128 hidden size to predict the value at $t = i + h$.

As an additional point of comparison, we also explore the text-only performance for each classification task. Since all datasets contain ground truth speech transcription, we take these transcriptions as the input data. As the NLP model we use the pre-trained RoBERTa [33] for SarD and SA, and Longformer [34] for PP[5]. We follow the typical method to fine-tune the pre-trained NLP model: extract the sentence embedding from the [CLS] token's embedding, and feed it to a linear classification model. No parameters are frozen during fine-tuning. We vary the learning rate ($[1e^{-3}$, $1e^{-4}$, $1e^{-5}$, and $1e^{-6}]$) and report the best performance.

4. MAIN RESULTS

4.1. SSL models perform well on prosody-intensive tasks

The experimental results for SA, SarD, and PP are shown in Figure 2. In Figure 2 (a) and (b), we can see that all SSL models yield better SA performance than the baseline FBANK features in both 2-label and 7-label evaluations. The large SSL models, wav2vec 2.0, HuBERT, and WavLM, even improve on the state-of-the-art (SOTA) performance [26] in the 2-label setup. In the 7-label setup, WavLM Large outperforms audio-only SOTA performance. Around one-third of the SSL models show better performance than the text-only baseline, confirming the value of acoustic information for SA.

Figure 2 (c) shows the SarD result. Although Mockingjay, wav2vec, and vq-wav2vec show inferior performance to the baseline FBANK, the other SSL models do well, with DistilHuBERT, HuBERT, and WavLM improving on the previous audio-only SOTA [30]. In SarD, all acoustic models, both SSL models and FBANK, outperform the text-only baseline, confirming that SarD requires acoustic information beyond content for prediction.

Lastly, for the PP results, Figure 2 (d), shows that all SSL models yield better performance than the FBANK features and the previous audio-only SOTA [31]. APC, VQ-APC, vq-wav2vec, and HuBERT obtains superior or equal accuracy to the text-only baseline.

[5]Longformer is based on RoBERTa, but can accept up to 4096 tokens (versus 512 tokens for RoBERTa), and is used here because the transcriptions of PP are too long for RoBERTa.

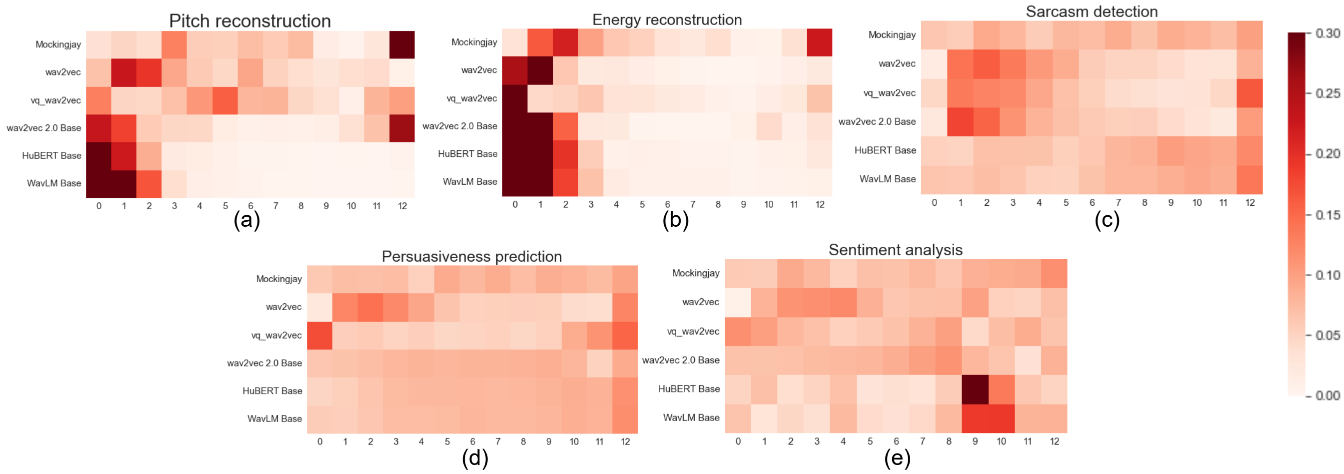

Fig. 3. The contribution analysis for each task. The darker the color, the higher the contribution. We only include models which have 12 layers of representations.

Model	Pitch w/ Prediction Horizon (s)↓				Energy w/ Prediction Horizon (s)↓			
	0.12	0.24	0.50	1.00	0.12	0.24	0.50	1.00
FBANK + RNN	0.049	0.104	0.142	0.157	0.52	1.42	2.06	2.46
APC	0.033	0.043	0.052	0.053	0.91	1.46	1.98	2.37
modified CPC	0.038	0.051	0.062	0.065	0.79	1.37	1.94	2.47
wav2vec	0.053	0.064	0.075	0.075	0.68	1.15	1.70	2.31
Mockingjay	0.069	0.077	0.081	0.099	0.48	1.92	2.17	2.43
wav2vec 2.0 Base	0.038	0.047	0.047	0.049	0.44	0.85	1.24	1.49
wav2vec 2.0 Large	0.035	0.039	0.045	0.046	0.43	0.80	1.23	**1.31**
HuBERT Base	0.029	0.036	0.041	0.042	**0.39**	**0.70**	**1.12**	1.42
HuBERT Large	**0.025**	**0.027**	**0.028**	**0.037**	0.41	0.77	1.21	1.36

Table 2. The MSE loss for FVP. Lower values indicate better prediction, of future pitch or energy.

4.2. SSL models encode prosodic information

While the results above suggest that the SSL models are encoding prosodic information, there is also direct evidence from the ProR and FVP tasks, as seen in Figure 2 and Table 2.

For PR, the two features, pitch and energy, have slightly different results. In pitch reconstruction, Figure 2 (e), all the SSL models perform better than baseline FBANK except for vq-wav2vec. Several SSL models surpass REAPER performance, with the WavLM the best. As for energy reconstruction, Figure 2 (f) shows that all SSL models greatly improved over baseline FBANK. Although generation-based SSL models perform well on pitch reconstruction, they did relatively worse on energy reconstruction. On the other hand, the SSL models pre-trained by masked contrastive discrimination/token prediction show strong performance on both pitch and energy reconstruction. Overall, we observe that SSL models indeed encode prosodic information.

FVP is more challenging than ProR since it requires the model to capture both global and local prosodic information for successful future prediction. From Table 2, we see, unsurprisingly, that the larger prediction horizons make prediction harder. For pitch, HuBERT Large gets the best performance, outperforming other SSL models and baseline FBANK+RNN at all four horizons. Although the pre-training objectives of APC, modified CPC, and wav2vec involve future generation/discrimination, they still result in inferior performance to wav2vec 2.0 and HuBERT. As for future energy prediction, only wav2vec 2.0 and HuBERT consistently outperform baseline FBANK. In general, we observe that some SSL models are not good at FVP, but wav2vec 2.0 and HuBERT outperform FBANK by a large margin. This result suggests that wav2vec 2.0 and HuBERT do have the capability to encode and summarize relevant prosodic information.

5. FURTHER ANALYSIS

5.1. Layerwise contribution analysis

In order to estimate the contribution of each layer, we consider two factors. First, we use the weight w_i from each layer (through the weighted-sum mechanism) to the downstream model. Second, because typical feature magnitudes may vary, we also consider the values of the hidden representation $\mathbf{x}$ in

Layer selection	SarD F1↑ (0,1,12)	SarD F1↑ (10,11,12)	PP Acc↑ (0,1,12)	PP Acc↑ (10,11,12)	SA Acc↑ (0,1,8)	SA Acc↑ (7,8,9)
wav2vec 2.0 Base	**70.6**	66.3	**81.3**	81.3	73.6	**74.0**
HuBERT Base	**72.0**	70.0	**85.8**	83.8	75.2	**76.2**

Table 3. Experimental results for layer-limited feature integration for SarD, PP, and SA. The best layer, underlined, is chosen based on the contribution analysis.

Method	ZH-p↓	PO-p↓	ZH-e↓	PO-e↓
FBANK	0.050	0.096	0.849	0.477
vq-wav2vec	0.022	0.097	0.490	0.339
wav2vec 2.0 Base	0.012	0.039	0.338	0.160
HuBERT Base	0.009	0.042	**0.232**	**0.149**
WavLM Base	**0.008**	**0.019**	0.233	0.192

Table 4. MSE for prosody reconstruction, showing cross-lingual transferability. p = pitch, e = energy.

each layer, as measured by the L2-norm of feature values for the testing data. After getting the whole L2-norm features for each layer for each sample, we take the mean across samples to get the feature magnitude estimation. The contribution for each layer is then defined as: $c_i = ||\mathbf{x_i}||_2 \times w_i$, where i means the layer number.

The results are shown in Figure 3. From Figure 3(a), we can see that for most SSL models, the contribution is strongest in the first few layers for both pitch and energy reconstruction. This shows that SSL tends to best represent prosodic information in the front. One exception is Mockingjay, where the largest contribution is located in the last layer. Because Mockingjay's pre-training objective is to reconstruct frame-wise features, it is unsurprising that the later layers contain a good representation of low-level prosody.

However, for the classification tasks SarD and PP, Figure 3 (c) and (d) shows that the distribution of layer contributions is smooth, suggesting that both tasks need information across multiple layers. As previous work has shown that later layers represent more content information [11, 25], this suggests that both prosodic and content information are needed for SarD and PP. As for SA, in Figure 3(e), we observe a high contribution value in the latter layers for the high-performing models HuBERT and WavLM. This suggests that SA might only require content information to perform well.

5.2. Feature integration

To further examine whether the encodings in the first few layers bring the most benefit, we designed a new experiment for three prosody-intensive tasks using wav2vec 2.0 and HuBERT. We compare two settings, concatenation of 1) the features from the first two layers and the best layer we discovered[6], and 2) the best layer with its two neighbor layers. These concatenated features are passed to the downstream model. The downstream model size of the two settings is the same, so we can compare the performance fairly. If the first setting is better than the second setting, this would indicate that early-layer (low-level) information indeed most benefits the final results.

The results are shown in Table 3. In SarD and PP, we see the integration of the first two layers yields better performance, which means the low-level information improves the modeling of SarD and PP. Yet the use of low-level information does not improve SA performance, suggesting that SA may rely on content rather than just low-level information.

[6]The best layer is determined by the contribution analysis above.

5.3. Cross-lingual transferability

Further, we did a preliminary investigation of whether SSL models may have cross-lingual transferability for prosodic information, using the ProR task as in previous experiments: specifically, attempting pitch and energy reconstruction. While the pitch of one frame is primarily a physical phenomenon, in context there may be language dependencies. All the SSL models being pre-trained in English, we try this for Mandarin and Polish, using data from AISHELL-3 [35] multilingual LibriSpeech (MLS) [36], respectively. Table 4 summarizes the results. We note that WavLM is good at pitch reconstruction, and HuBERT is superior at energy reconstruction.

6. CONCLUSION

In this work, we explore the utility of SSL models for prosody-conveyed pragmatic functions. Using the newly-proposed SUPERB-prosody evaluation framework, we find that SSL models provide significant value for prosody-intensive tasks, and that they are good at extracting prosodic information in pseudo tasks. Furthermore, we analyze the layer contribution and discover that most SSL models tend to store prosodic information in the first few layers. However, the field still lacks a good understanding of why different SSL models are better for different tasks [37], and this is an important topic for future work.

Acknowledgement: This work was supported by JSALT 2022 at JHU, with gift-funds from Amazon, Microsoft, and Google. We thank Taiwan Web Service for providing computational and storage resources.

7. REFERENCES

[1] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[2] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, et al., "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8479–8492.

[3] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[4] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, et al., "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, pp. 101255, 2022.

[5] Nigel G. Ward, *Prosodic Patterns in English Conversation*, Cambridge University Press, 2019.

[6] Oliver Niebuhr, "Stepped intonation contours: A new field of complexity," in *Tackling the Complexity of Speech*, R. Skarnitzl and O. Niebuhr, Eds., pp. 39–74. Charles University Press, 2015.

[7] Jack Weston, Raphael Lenain, Udeepa Meepegama, and Emil Fristed, "Learning de-identified representations of prosody from raw audio," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 11134–11145, PMLR.

[8] Diego Aguirre, Nigel Ward, Jonathan E Avila, and Heike Lehnert-LeHouillier, "Comparison of models for detecting off-putting speaking styles," *Proc. Interspeech*, pp. 2303–2307, 2022.

[9] Joel Shor, Aren Jansen, Ronnie Zvi Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, and Dotan Emanuel, "Towards learning a universal non-semantic representation of speech," in *Proc. Interspeech*, 2020.

[10] Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han, "SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7927–7931.

[11] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[12] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[13] Yu-An Chung, Yonatan Belinkov, and James Glass, "Similarity analysis of self-supervised speech representations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3040–3044.

[14] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Interspeech*, 2019, pp. 146–150.

[15] Yu-An Chung, Hao Tang, and James Glass, "Vector-quantized autoregressive predictive coding," in *Interspeech*, 2020, pp. 3760–3764.

[16] Alexander H Liu, Yu-An Chung, and James Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv preprint arXiv:2011.00406*, 2020.

[17] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[18] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[19] Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020, pp. 7414–7418.

[20] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition.," in *Interspeech*, 2019.

[21] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[22] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[23] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

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

[25] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[26] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2236–2246, Association for Computational Linguistics.

[27] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria, "Towards multimodal sarcasm detection (an _Obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 4619–4629, Association for Computational Linguistics.

[28] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency, "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 50–57.

[29] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *CoRR*, vol. abs/1904.02882, 2019.

[30] Ning Ding, Sheng-wei Tian, and Long Yu, "A multimodal fusion method for sarcasm detection based on late fusion," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8597–8616, 2022.

[31] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[32] David Talkin, "Reaper: Robust epoch and pitch estimator," https://github.com/google/REAPER, 2015.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[34] Iz Beltagy, Matthew E Peters, and Arman Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[35] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker Mandarin TTS corpus and the baselines," 2015.

[36] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[37] Sri Harsha Dumpala, Chandramouli Shama Sastry, Rudolf Uher, and Sageev Oore, "On Combining Global and Localized Self-Supervised Models of Speech," in *Proc. Interspeech 2022*, 2022, pp. 3593–3597.