Multimodal Prosody Modeling: A Use Case for Multilingual Sentence Mode Prediction

Bogdan Vlasenko, Mathew Magimai.-Doss

[1]Idiap Research Institute, Martigny, Switzerland

bogdan.vlasenko@idiap.ch, mathew@idiap.ch

Abstract

Prosody modeling has garnered significant attention from the speech processing community. Recent developments in multilingual latent spaces for representing linguistic and acoustic information have become a new trend in various research directions. Therefore, we decided to evaluate the ability of multilingual acoustic neural embeddings and knowledge-based features to preserve sentence-mode-related information at the suprasegmental level. For linguistic information modeling, we selected neural embeddings based on word- and phoneme-level latent space representations. The experimental study was conducted using Italian, French, and German audiobook recordings, as well as emotional speech samples from EMO-DB. Both intra- and inter-language experimental protocols were used to assess classification performance for uni- and multimodal (early fusion approach) features. For comparison, we used a sentence mode prediction system built on top of automatically generated WHISPER-based transcripts.

Index Terms: multimodal, multilingual, sentence mode prediction, emotional prosody

1. Introduction

Prosody plays an important role in human-to-human communication. Prosodic cues are used to define boundaries between linguistic units, specify the accentuation of a sentence, or indicate different types of sentence moods. In addition to the linguistic aspects of prosody, it can also be used to add emotional cues to speech. Hence, proper prosody modeling is an important research direction in speech and natural language processing. Each language has unique prosodic patterns, making prosody a language-specific phenomenon from a linguistic perspective. The first attempt to establish a standardized prosodic transcription system for English was introduced in [1], where the authors developed the Tones and Break Indices (TOBI) framework for prosodic event transcription. Later, the German TOBI (GToBI) [2] standard was created to enhance prosody modeling in German. Linguists continue to refine and develop TOBI-like standards for various languages. However, TOBI standards are language-specific, posing challenges for multilingual prosody modeling. Additionally, generating accurate TOBI-based prosodic transcriptions is costly and requires expert linguists for speech annotation. One of the earliest research efforts in multilingual prosody modeling can be traced back to the Verbmobil project [3]. The Verbmobil project, addressed the importance of prosody modeling for advanced speech understanding techniques. In [4], project contributors described a speech-to-speech translation system—the first complete system to successfully integrate prosodic information into linguistic analysis. They demonstrated how prosody could be leveraged to compute probabilities for clause boundaries, accentuation, and sentence mood classification. The project's primary goal was to enable automatic speech translation by transferring prosodic patterns across selected languages. Additionally, Verbmobil's partners emphasized the importance of prosody modeling for the development of text-to-speech (TTS) [5] models.

Early attempts at prosody modeling in TTS systems relied on knowledge-based features. Studies have shown that F0 modeling [6–8] can enhance the prosodic characteristics of synthesized speech. With the advent of self-supervised learning (SSL) [9, 10] techniques and data-driven approaches in TTS [11], prosody modeling has evolved to incorporate advanced linguistic and acoustic neural embeddings generated by various pre-trained SSL models [12, 13]. Recent findings in voice conversion techniques have shown that in the FreeVC [14] approach, WavLM [15] feature representations (FRS), processed through a bottleneck extractor, serve as encoder information for the VITS [16] TTS system. In [17], researchers employed an emotion-based encoder built on the pre-trained Wav2Vec2-XLSR [18] model to enhance expressivity transfer in text-less speech-to-speech translation. To make prosody modeling applicable to multilingual setups, several research groups have proposed using phone-level BERT embeddings [19, 20], modeled on linguistic information from large multilingual text corpora.

Recent developments in end-to-end Automatic Speech Recognition (ASR) systems [21, 22] have demonstrated that, with sufficiently large models, joint multilingual and multi-task training offers no drawbacks—and even provides benefits. In addition to a multilingual training and processing pipeline, WHISPER-based ASR [21] systems enable deep integration of modality modeling (linguistic—language modeling; acoustic—acoustic models) within an advanced neural architecture. This technique is based on a large-scale weak supervision concept applied to extensive speech datasets with partial transcription (textual transcripts without proper prosodic annotations). As noted earlier, standardized speech transcription with prosodic characteristics is a complex, language-specific process requiring professional linguists. Therefore, in our study, we focused on providing sentence mode prediction for multilingual settings. For the data-driven FRS representations we used pre-trained *multilingual* SSL models to investigate multilinguality from the perspective of a common for all modeled languages latent space for both modalities: acoustic and linguistic. In our study, we are addressing the following research question:

1. How well can sentence mode be predicted using uni-modal (acoustic or linguistic) and multimodal processing techniques?

2. How well can sentence mode prediction performance be achieved in cross-lingual experimental settings?

3. How well do sentence mode prediction techniques perform on emotional speech?

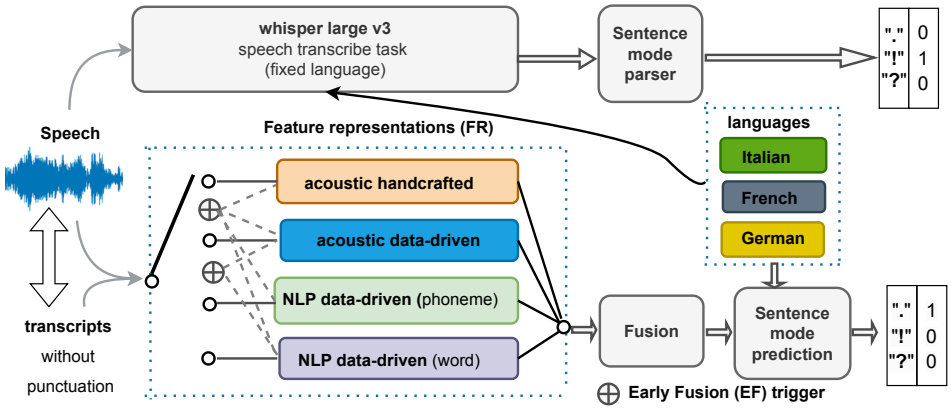


Figure 1: *Overview of the proposed experimental study*

2. Methodology

This section presents our methodology and datasets used in our study. The processing flow of our multimodal sentence mode prediction is presented in Figure 1. In our study we employed acoustic and linguistic suprasegmental FRs which shoved outstanding classification performance for speech based emotion recognition. Acoustic FRs were represented using both handcrafted and data-driven features, while linguistic FRs were data-driven with different levels of representation: words and phonemes. For the data-driven FRs, we employed neural embeddings extracted from the final layer of a pre-trained neural model designed for general speech processing tasks. In the case of data-driven acoustic and linguistic FRs, the sequence of frame and word/phoneme-level embeddings was mapped into a fixed-length vector using the mean function. In the case of knowledge-based FRs, we used a broader pool of functionals to generate fixed-length vectors for our suprasegmental modeling. To remove prior information about punctuation, we removed punctuation marks before extracting the linguistic FRs for the entire sentence. To evaluate multimodal aspects, we used an early fusion (EF) technique applied to fixed-length feature representations. Two configurations of EF were used: a data-driven acoustic + data-driven linguistic (sentence-level) setup and a combined setup for all acoustic and linguistic FRs. For classification purposes we employed MLP classifier. In order to compare our proposed multimodal and multilingual sentence mode prediction systems, we created a sentence mode prediction system based on transcripts obtained with pre-trained multilingual Automatic Speech Recognition (ASR) system - WHISPER LARGE V3. More technical details on implementation side could be found in Section 3.

Datasets: In the proposed study we used corpora with multilingual audiobook recordings - CML-TTS [23] and auxiliary emotional corpus EMO-DB [24] with acted German emotional speech samples. We modeled three different sentence modes: *exclamatory* (punctuation mark - **!**), *interrogative* (punctuation mark - **?**), and a combined class of *declarative* sentences (punctuation mark - **.**). In our study, we used speech samples from three languages: Italian, French, and German. Although sentence modes can be defined differently across languages, in our study the three punctuation marks, namely, **[“.”]**, **[“!”]**, **[“?”]**) are treated in a similar way across the selected languages.

Table 1: *Number of samples per sentence mode.*

sign/lang.	Italian	French	German
[“.”]	2870	4043	19346
[“!”]	1708	1751	14113
[“?”]	1648	1366	12259
Total	**6226**	**7160**	**45718**

For the CML-TTS database, we wrote pre-selection script to generate labeled samples for sentence mode prediction task. Originally, the CML-TTS database provided transcripts with punctuation applicable for training text-to-speech models. The pre-selection script selected and assigned classification labels based on the final punctuation mark. The number of speech samples for each sentence mode can be found in the table 1. To make our study easier to reproduce, the list of speech samples (Italian, French, and German) with corresponding sentence mode labels can be found in our GitHub repository.[1] For the emotional speech corpora, we selected acted emotional speech datasets with a predefined list of sentences. The EMO-DB database includes 9 declarative sentences ending with a period “.” and 1 interrogative sentence ending with a question mark “?”. In our experimental study, we used all 535 speech samples from the EMO-DB: 79 neutral and 456 emotional. For our experimental study, we use acoustic signals from wave files presented in selected datasets, along with corresponding manual transcripts with punctuation removed (see Figure 1).

3. Experimental setup

The section introduces system designs with corresponding experimental protocols, evaluation metrics, implementation tools used for the study.

Systems: The *first* system in our study used all the previously mentioned uni-modal FRs and two EF concepts described earlier. To evaluate the intra-language experimental protocol, we used a 5-fold cross-validation technique. For the inter-language protocol, we trained classifiers on one language and evaluated the obtained models on another.

For the *second* system, we used the WHISPER [2] ASR model for transcription tasks with a predefined language. We then applied sentence mode detection based on punctuation markings in the generated transcription texts. If both ’?’ and ‘!’ appeared in the same transcript of a wave file, we assumed that second system made a mistake. A list of wave files with double punctuation marks will be provided in our GitHub repository.

The *third* system in our study was used for the intra-

[1] https://github.com/idiap/IS2025_MPM
[2] https://huggingface.co/openai/whisper-large-v3

Table 2: *Sentence mode prediction performance in* UAR [%] *for system 1. Intra- and inter-language setups,* CML-TTS *database.*

				EVALUATION LANGUAGE					
				Italian		**French**		**German**	
Training	**Type FR**	**Version I**	**Version II**	UAR I	UAR II	UAR I	UAR II	UAR I	UAR II
Italian	ACOUSTIC	*knowledge*	*data-driven*	51.51	55.71	41.27	48.67	43.45	48.30
	LINGUISTIC	*phone*	*word*	53.76	58.85	44.84	56.60	41.42	64.54
	EARLY FUSION	*data-driven*	*all*	**64.14**	59.40	57.42	45.31	59.38	49.87
French	ACOUSTIC	*knowledge*	*data-driven*	41.38	47.61	46.28	55.56	41.50	51.21
	LINGUISTIC	*phone*	*word*	45.79	47.03	52.20	60.14	44.87	57.56
	EARLY FUSION	*data-driven*	*all*	50.52	47.68	**67.10**	63.22	63.13	53.42
German	ACOUSTIC	*knowledge*	*data-driven*	44.61	47.78	45.63	49.47	56.01	61.53
	LINGUISTIC	*phone*	*word*	39.55	57.54	39.24	62.15	60.04	69.17
	EARLY FUSION	*data-driven*	*all*	54.45	58.51	60.75	63.62	**74.22**	73.40

language and cross-corpora experimental protocols. We trained a sentence-mode classifier on the CML-TTS German subset and evaluated the obtained models on EMO-DB database samples. Due to the fixed linguistic content of EMO-DB (10 predefined sentences), we used only acoustic FRS

Metrics: Considering the imbalance in the number of instances per class in the selected speech corpora, we decided to use unweighted-average recall (UAR) to measure sentence mode prediction performance and class-level recalls to measure classification performance per each sentence mode.

Implementation tools: In order to generate selected FRS we used publicly available tools and pre-trained models. Descriptions and links for the pre-trained models used for FRS generation were downloaded from Hugging Face, and those used for FRS extraction can be found in our GitHub repository.

Acoustic FRs: For the *knowledge-based* handcrafted FRS, we use COMPARE 2016 [25]. Feature set contains 6373 static turn-level features. were The python OPENSMILE package [26] was used to extract our knowledge-based FRS.

Considering top performance positions on challenge leaderboards for self-supervised-learning SSL embeddings [27] of general purposed for various tasks in SUPERB challenge [28] and SUPERB-prosody challange [29] we employed WAVLM (large) [15] embedding.

Linguistic FRs: For linguistics-based modeling, we employed embeddings extracted at the phoneme and word-levels. In figure 1, we refer to these features as NLP data-driven (phoneme) and NLP data-driven (word), respectively.

For *word-level* embedding representation, we used the XLM-ROBERTA [30] model, pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages. Furthermore, the selection of phoneme-level FRS was justified by the strong performance of emotional prosody modeling based on phoneme-level acoustic features [31, 32]

Using *phoneme-level* latent space based on textual information processing has become a recent trend in the TTS field [33]. In our study, we employed XPHONEBERT [19].

Table 3: *Recognition performance of system 2.* WHISPER *ASR.*

LANGUAGE	UAR[%]	RECALL [%]		
		[“.”]	[“!”]	[“?”]
Italian	57.03	96.72	20.55	53.82
French	64.10	97.40	15.25	79.65
German	58.99	98.75	14.55	63.66

Classifier: The MLP classifier with 100 neurons in a single hidden layer and RELU activation function were used for training sentence mode prediction models. MLP was trained with sklearn library in python.

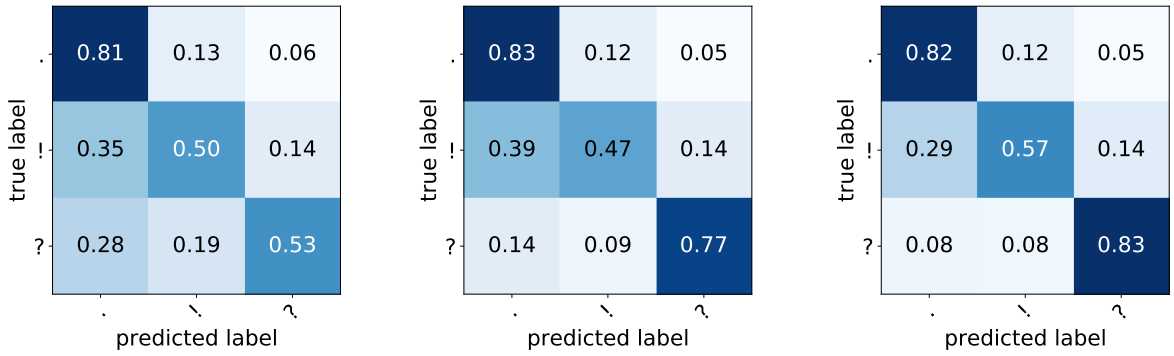
4. Results

Table 2 presents UAR rates obtained with first system during intra- and inter-language experimental setups. The results for the intra-language experimental protocol can be found in the diagonal of the table (i.e., matched conditions for training and evaluation languages). The remaining (non-diagonal) result blocks correspond to the inter-language experimental protocols. Results for the best-performing FR combinations are highlighted in bold for intra-language settings and underlined for inter-language experimental protocols. For each type of FR we used two versions: knowledge-based vs. data-driven (acoustic); phoneme- and word-level for linguistic data-driven FRS; data-driven (acoustic and linguistic word-level) vs. all (four types of acoustic and linguistic FRS). Results UAR I corresponds to Version I and UAR II represent recognition rates for Version II.

Table 3 presents sentence mode prediction performance for second system, based on WHISPER ASR model. Class-wise recall rates for exclamatory sentence ([“!”]) are lower then selection by chance rates (33%) for all evaluated languages. List of miss-recognized sentences (double punctuation signs ([“!”] and [“?”])) for the 2nd system will be presented in our GitHub site. Finally classification performance for the third system presented in Table 4. The table contains class-level recall rates obtained during the cross-corpora sentence mode prediction study. MLP models trained (CML-TTS German subset) of three types of sentence modes in German language we evaluated on EMO-DB database samples which contain instances for combined class declarative sentences (punctuation mark [“.”]) vs. interrogative sentences (punctuation mark - [“?”]).

Table 4: *Class-wise recall [%] for system 3.* EMO-DB *dataset.*

TYPE OF SPEECH	ACOUSTIC FR			
	knowledge		*data-driven*	
	[“.”]	[“?”]	[“.”]	[“?”]
NEUTRAL	9.72	42.86	**52.78**	**42.86**
EMOTIONAL	2.18	43.18	13.60	20.45



(a) *Acoustic data-driven* (b) *Linguistic data-driven word-level* (c) *EF data-driven*

Figure 2: *Confusion matrices for sentence mode classification. Intra-language evaluation for German language samples from* CML-TTS *database. Abbreviations:* EF - *early fusion.*

5. Analysis

Results in Table 2 show that linguistic and acoustic neural embeddings, trained on large multilingual corpora, can be used as FRs for effective sentence mode classification in cross-language settings. It can be observed that the most discriminative features for intra- and cross-language experimental protocols are obtained through the early fusion of data-driven word-level embeddings (XLM-RoBERTa) and data-driven acoustic neural embeddings (WavLM). Impressive recognition rates were observed in a cross-corpora setup, where models were trained on German speech and evaluated on French speech. Models trained on word-level linguistic embeddings from German (UAR=62.15%) outperformed the results obtained in an intra-corpora setup (UAR=60.14%). For sentence-mode predictive models trained on German speech samples and evaluated on Italian and French sentences, the baseline inter-language UAR rates were obtained using early fusion of *all* FRs. The results from early fusion outperform the performance achieved with the *data-driven EF* setup: train German - test Italian: 54.45% vs. 58.51; train German - test French: 60.75% vs. 63.62%. Hence, we assume that using phoneme-level BERT combined with knowledge-based acoustic FRs could provide beneficial information for cross-lingual settings, especially when the experimental languages belong to different language groups (Germanic and Romance languages).

Figure 2 presents confusion matrices for an intra-language experimental study conducted on German speech samples from the CML-TTS database. As seen in Figure 2a and Figure 2b, acoustic FRs provides slightly better class-wise recall rates for exclamatory sentences. On the other hand, word-level linguistic FRs has significantly better class-wise recall rates for interrogative sentences. Finally, as shown in Figure 2c, applying early fusion to data-driven acoustic and word-level linguistic FRs can boost class-wise recall rates for both sentence modes: interrogative and exclamatory. The complete pool of CF matrices can be found in our GitHub's page.

Table 4 results show that sentence mode prediction models trained on acoustic data-driven FRs yield recall rates (53% and 43%) above chance level (33.33%) on neutral speech from the EMO-DB database. On the other hand for emotional speech samples sentence mode prediction dropped significantly for both types of acoustic FRs. Also, most of the emotional speech samples presented in EMO-DB database were miss-recognized as exclamatory sentences (punctuation mark - ["!"]).

6. Conclusion

Addressing our first research question, we found that linguistic FRs contain more sentence-mode-related information compared to acoustic FRs. In our study, we demonstrated that linguistic embeddings serve as a valuable source of sentence-mode-specific, prosody-related information. Our findings align with recent research [22, 34], which explores the enhancement of prosodic characteristics using SSL-based embeddings as encoder representations. On the other hand, early fusion of data-driven linguistic features (word-level embeddings) and acoustic turn-level multilingual FRs provides the best performance for intra-language experimental protocols, even though the evaluated feature representation primarily reflects simple suprasegmental features.

Addressing our second research question, we showed that data-driven FRs extracted from a common multilingual latent space can be used for reliable cross-language sentence-mode prediction. In cross-lingual sentence-mode prediction experiments (models trained on German speech samples), we demonstrated that the fusion of all FRs (all four types: data-driven + knowledge-based) can improve sentence-mode prediction. This suggests that, in some cross-lingual experimental settings, phoneme-level BERT and knowledge-based acoustic FRs can enhance cross-lingual sentence-mode prediction performance.

In the case of the third research question related to sentence-mode prediction on emotional speech samples, we showed that sentence-mode prediction performance deteriorates on emotional speech samples. Sentence-mode prediction models trained on German audiobook speech samples showed RECALL["?"]=42.86 and RECALL["."]=52.78 for neutral speech from EMO-DB, and RECALL["?"]=20.45 and RECALL["."]=13.60 for emotional speech from EMO-DB. The definition of exclamatory sentences as 'sentences that convey strong emotions or feelings' was observed in our cross-corpora experimental study. Hence, most of the German emotional speech samples with a single non-exclamatory tone were misclassified as exclamatory by sentence-mode prediction models trained on CML-TTS German samples.

Experimental results obtained with the second system, based on the Whisper ASR transcripts, showed that recent state-of-the-art multilingual ASR models are unable to detect exclamatory sentences, even though the selected speech data consists of well-articulated audiobook recordings. On the other hand, the class-wise recall rates for declarative and exclamatory sentences were high.

7. Acknowledgments

This work was partially funded by the Innosuisse through the flagship project IICT: Inclusive Information and Communication Technologies (grant agreement no. PFFS-21-47).

8. References

[1] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody." in *ICSLP*, vol. 2, 1992, pp. 867–870.

[2] M. Grice, M. Reyelt, R. Benzmuller, J. Mayer, and A. Batliner, "Consistency in transcription and labelling of German intonation with GToBI," in *Proc. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1716–1719.

[3] W. Wahlster, *Verbmobil: Translation of face-to-face dialogs*. Springer, 1993.

[4] E. Noth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Transactions on Speech and Audio processing*, vol. 8, no. 5, pp. 519–532, 2000.

[5] W. Wahlster, "Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final verbmobil system," *Verbmobil: Foundations of speech-to-speech translation*, pp. 3–21, 2000.

[6] S. Latif, I. Kim, I. Calapodescu, and L. Besacier, "Controlling prosody in end-to-end tts: A case study on contrastive focus generation," in *Proc. CoNLL*, 2021, pp. 544–551.

[7] J. P. Teixeira *et al.*, "A prosody model to tts systems," *Phd, Faculdade de Engenharia da Universidade do Porto (May 2004)*, 2004.

[8] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: considerations for tts," in *Speech Prosody 2008*, 2008, pp. 661–665.

[9] T. Purohit, B. Vlasenko, and M. Magimai.-Doss, "Implicit phonetic information modeling for speech emotion recognition," in *Proc. Interspeech 2023*.

[10] B. Vlasenko, S. Vyas, and M. Magimai.-Doss, "Comparing data-driven and handcrafted features for dimensional emotion recognition," in *Proc. ICASSP 2024*.

[11] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, and J. Yamagishi, "Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4036–4051, 2024.

[12] A. R. Naini, M. A. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *Proc. ICASSP 2024*, pp. 12 031–12 035.

[13] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[14] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP 2023*.

[15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[16] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[17] J. Duret, B. O'Brien, Y. Estève, and T. Parcollet, "Enhancing expressivity transfer in textless speech-to-speech translation," in *Proc. ASRU 2023*, 2023.

[18] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[19] L. T. Nguyen, T. Pham, and D. Q. Nguyen, "XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech," in *Proc. Interspeech 2023*.

[20] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions," in *Proc. ICASSP 2023*.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[22] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[23] F. S. Oliveira, E. Casanova, A. C. Junior, A. S. Soares, and A. R. Galvão Filho, "Cml-tts: A multilingual dataset for speech synthesis in low-resource languages," in *Proc. TSD 2023*. Springer, pp. 188–199.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[25] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE — The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACMM MM*, 2010, pp. 1459–1462.

[27] S. Yadav, T. Purohit, Z. Mostaani, B. Vlasenko, and M. Magimai.-Doss, "Comparing biosignal and acoustic feature representation for continuous emotion recognition," in *Proc. of MUSE ACM MM 2022*.

[28] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[29] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *Proc. SLT 2022*, pp. 1104–1111.

[30] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, 2019.

[31] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Proc. ACII 2013*, Geneva, Switzerland, September 2013, pp. 511–516.

[32] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483–500, 2014.

[33] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions," in *Proc. ICASSP 2023*.

[34] L. Chen, Y. Deng, X. Wang, F. K. Soong, and L. He, "Speech BERT Embedding for Improving Prosody in Neural TTS," in *Pros. of ICASSP*, 2021, pp. 6563–6567.