Prepublication version of paper to appear in Speech Prosody 2026

ML Classification in a Benchmark of Prosodic Minimal Pairs

Sahya Lagisetty[1], *Mats Rooth*[2]

[1] Highland Park High School, Dallas TX
[2]Cornell University

sahya.lagisetty@gmail.com, mr249@cornell.edu

Abstract

This paper presents the following: (i) A computational methodology for collecting large numbers of utterances of a fixed word string from online sources, using the index of `youtube` transcriptions at `filmot.com`. (ii) A prototype benchmark constructed with the methodology consisting of *prosodic minimal pairs*, which are short word sequences which depending on context and/or lexical identity are pronounced with different prosodies. The benchmark is grouped into pairs, for instance utterances of "much as I did" with or without focus prosody on the first person subject. (iii) A classification model obtained by tuning wav2vec2-base on the training portion of the benchmark. Presented with an audio that is stipulated to be an utterance of a given word string, the model selects one of two alternative prosodies for the word string. For instance, it can determine whether a given utterance of "much as I did" has focus prosody on the subject, or default prosody. Accuracy of this determination on separate test data is better than 90% for each minimal pair.

Index Terms: prosody, audio classification, alternative focus, stress doublet

1. Introduction

Trying to confirm or refute predictions of linguistic theories of prosody in large bodies of corpus data is an old vision. In the area of focus prosody, a precursor to the work here is research reported in Jonathan Howell's dissertation and related publications [1, 2, 3], where about two hundred utterances of the string "than I did" were collected from online radio stations that had speech recognition indexing. Transcriptions were corrected by experimenters, and classified by experimenters according to a criterion of varying reference in the subject position of a comparative than-clause, predicting focus prosody, versus constant reference in the subject position (relative to the main clause), predicting default prosody. Howell labeled phone boundaries in Praat using acoustic landmarks, and computed dozens of measures including phone durations, pitch, spectral tilt, and formant spread. SVM and LDA models were estimated for subsets of the features, and evaluated on held-out data. Models based on some sets of features achieved classification accuracy approaching 90% as measured using cross-validation.

Promising as these methods and results are, they have not been taken further, for reasons that are conjectured to involve the large amount of human annotation required, the poor quality of transcriptions at the time, and the limited amounts of indexed data available at the time. It is possible to do better using current computational methodology and data sources. In particular, it is possible to get more data, and eliminate much of the human annotation. First, the index of Youtube maintained at `filmot.com` reports 580 thousand tokens of "than I did". Our experience indicates that circa a tenth of these can be expected to be retrievable. Thus experience suggests that currently it might be possible to obtain on the order of 10^4 or 10^5 tokens, rather than on the order of 10^2. Second, Youtube automatic transcripts are already good, and it is possible to rerun speech recognition and word alignment to get cleaner ones. Third, because so much data is available, longer search targets can be used to find tokens with a given prosody. For instance, the longer string "you enjoyed it as much as I did" conditions focus prosody on the subject in the minimal pair string "much as I did", because the contrasting subject "you" is in the long string. The string "as much as I did before" conditions non-focus prosody for the subject in "much as I did", because in practice, times and not agents are contrasted. Thus for some targets, it is possible to obtain tokens with identified prosody, without experimenter annotation. Fourth, models for audio classification are available that work from a low-level representation of the sound signal derived from a neural model, rather than measured features such as vowel duration and formant spread. Such models are tuned to the distinctions and labels found in the training portion of a benchmark, starting from a base model. They are successful in distinguishing sounds in many vocabularies of distinctions, or instance distinguishing different urban sounds. Using such models eliminates the need for explicit feature measurement, and the need to identify the sets of features that are relevant for a prosodic distinction.

2. Retrieval

`Filmot.com` is an index of speech recognition transcripts for videos at `Youtube.com`, where it is possible to search for words and word sequences. Results are presented to the users with embedded youtube video players, accompanied by clickable speech recognition transcripts coming from Youtube. See Figure 1. Information in a query including the target phrase is encoded in a URL. This makes it straightforward to retrieve `html` pages for hits with a software library which can retrieve the html page corresponding to a given URL. This was coded in several versions, initially using Unix shell and Awk, and in the final version using the Python pyCurl module.[1] We identified target phrases (see below) and retrieved HTML pages for up to 1000 hits per target or more, when available.

HTML pages having been retrieved, they are parsed to extract the video ID, the transcript in the area of the hit surrounding a token of the target phrase, and time locations for parts of the transcript. These are assembled in a tabular format.[2] The entire audio file for the hit is retrieved from youtube using `yt-dlp` [4] and this is cut to an approximately ten second clip that is expected to contain the hit, called the medium audio. The medium audio is re-transcribed into words with time align-

[1] See `script/retrieveCurl.py` in the code repository.

[2] See the sample data `data/the+present+you/id3_1000.tsv` in the code repository.

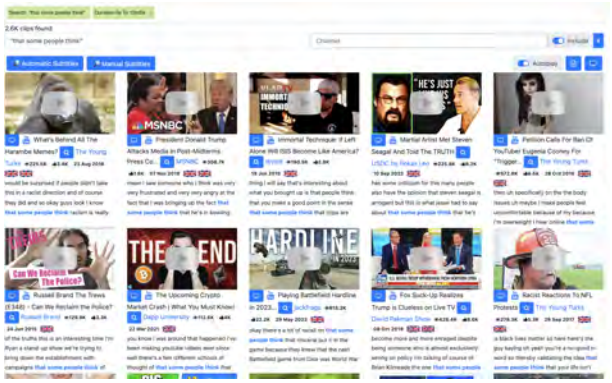

Figure 1: *Result display for a search for* that+some+people+think *at* `filmot.com`*, limited to videos of length at most ten minutes.*

ments using *stable-ts* [5, 6].[3] Finally the medium audio is cut to a short segment with an utterance of the minimal pair phrase, using the word time alignments from `stable-ts`. In the same stage, to allow for examination of examples, Praat textgrids for the medium audio with the transcription from `stable-ts` are created.

The following general strategy was followed. By informal observation, short target phrases such as "some people" that show variable prosody were identified. Through interaction with `filmot.com`, longer target phrases such as "that some people did" and "meet some people" were identified which in practice condition one prosody or the other, and which are sufficiently frequent. These were the search targets for the procedure above.

Due to the organization of `filmot.com`, for a given query, only a thousand hits can be retrieved. This is fixed by setting date ranges. In some cases, it is necessary to combine a number of longer target phrases triggering a given prosody, in order to obtain sufficiently many examples.

While the retrieval procedure is in principle straightforward, issues involving changed formats of web pages, restrictions on access from software, and updates of open source software for retrieval frequently come up.

3. The benchmark

The prototype benchmark consists of two minimal pairs involving focus prosody, and two stress doublets. In a notation with plus signs separating de-capitalized words, the two focus minimal pairs are *much+as+i+did*, and *some+people*.

The phrase *some people* frequently comes with a prominence on the determiner. This is theorized to correlate with a scalar implicature "some, but not all" [8]. In the longer phrase *but+some+people+think*, focus prominence on *some* is almost exceptionless. Sometimes though, *some+people* is pronounced with an unstressed determiner, and default prominence on *people*. This is almost exceptionlessly so in the larger phrase *meet+some+people*. See the examples (1).[4]

(1) a. let's get this all out there now there was reports in the english newspapers that some people think three lions would look very good on your chest
b. i don't know if i actually like yoga but i think now that this is like at my neighborhood spots i'll bet to meet some people in the area

The phrase *much+as+i+did* tends to come with focus prominence on the first person subject when the subject of the corresponding main clause does not refer to the speaker. This is the same correlation as the one exploited for *than+i+did* in [2], but for an equative clause rather than a comparative one. The long phrase *you+enjoyed+it+as+much+as+i+did* meets this criterion of varying reference, because a second person pronoun in the main clause is included in the search phrase. Tokens of the longer phrase *as+much+as+i+did+before* usually have constant subject reference, because a time and not an agent is contrasted. See the examples in (2).

(2) a. thank you so much for joining us today on the show i hope you enjoyed it as much as i did listening to the ui i hope you enjoyed it
b. and things like that i was a vegetarian for a very long time i eat meat now but not as much as i did before i like fish and i love shrimp and sushi

The above example types involve either focus prosody or contrastive topic prosody [9, 10]. All of them have a transparent alternative-semantics motivation for the prominence.[5]

"Many disyllabic words take stress on the first syllable when used as nouns and on the second when used as verbs" [11]. Such words are "stress doublets". Two are included in the benchmark, as counterpoints to the focus minimal pairs. Noun and verb uses can be found by specifying word contexts. In the phrase *my+object*, *object* is almost always a noun, with initials stress. In the phrase *i+object+to*, *object* is almost always a verb, with final stress. See the examples in (3). In the phrase *the+present+you*, *present* is almost always a noun, with initial stress. In the phrase *i+present+you*, it is almost always a verb, with final stress. See the examples in (4).

(3) a. section now of solving that slab problem and of course my line weights here are being controlled by my object styles so there's my section starting to
b. share my convictions and i think i'm the only candidate in this field that is a true free speech advocate to the core i object to governmental backdoor

[3] See `stablets.py` in the code repository. `Stable-ts` was run on CPU, which is fast enough. In an initial version of the workflow, the transcript from youtube was force aligned to the word level using Kaldi.[7]. In the final version of the workflow, the youtube transcript is used just to find the time interval for the medium audio, with the transcript and word timestamps coming from `stable-ts`.

[4] While these transcriptions are difficult to follow in written form, this largely reflects the complexity of spoken language, rather than errors in the transcriptions. What is said is usually easy to follow when listening to a ten second audio.

[5] An additional focus target *in+my+case* is under development, but is omitted here because the locus of focus turned out not to be consistent for half of the minimal pair. In the longer phrase *but+in+my+case*, the phrase *in+my+case* almost always is pronounced with prominence on the pronoun *my*. This comes with a discourse structure that explicitly or implicitly contrasts the speaker with others. In the longer phrase *involved+in+my+case*, we thought that the noun *case*, which usually refers to a legal or medical case, would come without prominence on *my*, and with default phrasal prominence on *case*. This turned out not to be consistent.

(i) when you click on this it's going to tell you to save it in the cloud but in my case i do have this one saved in the cloud

(ii) spent four hours there, only to find that the officer involved in my case called in sick or had to go home sick

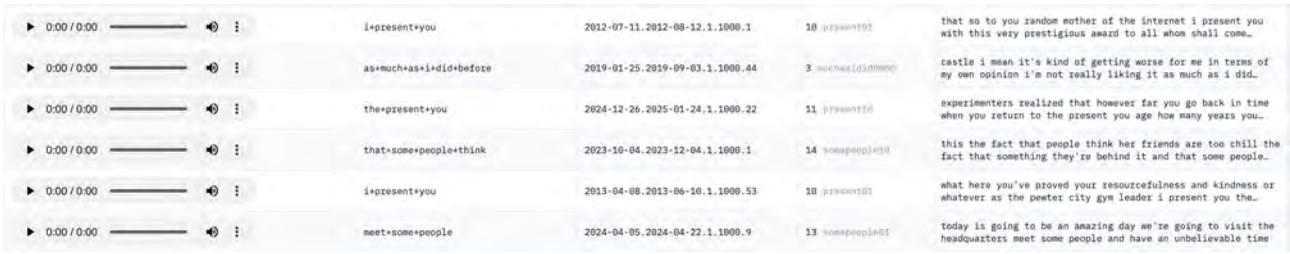

Figure 2: *An image from huggingface.co/datasets/MatsRooth of six items of the benchmark, with playable audio, long target, ID, label in numerical and text format, and text transcription. In the fourth line,* that+some+people+think *is the long search phrase that conditions focus prosody on* some *in the minimal pair phrase* some+people*. The corresponding label is* some+people10*. In the sixth line,* meet+some+people *is the search phrase that conditions default prosody in the minimal pair phrase* some+people*. The corresponding label is* some+people01*. Only the audio and the label (fourth column) are used in machine learning classification.*

(4) a. that first of all i have a strong belief that if you are going to invest in the stock market you do not invest in the present you do not pay attention to what
b. fear we must prove we are strong enough to handle our power responsibly you have done this i present you with your robes your staff and a ring bearing the

The stress doublets have differences in vowel quality as well as stress, /ˈɑbdʒɛkt/ vs /əbˈdʒɛkt/ and /ˈprɛzənt/ vs. /prɪˈzɛnt/ in standard transcriptions. This will make the classification task in the next section easier. However, also focus pairs are likely to have what amount to differences in vowel quality, due to different patterns of reduction and hyperarticulation.

The format of the benchmark adopts the open source Datasets library, which includes functionality for audio datasets [12]. To the basic attributes `audio` and `label` for audio datasets, we added attributes `target` for a long search string such as *that+some+people+think*, `ID` for an ID that us derived from our retrieval procedure, and `text` for the transcription of the ten second interval. See Figure 2. The attributes `target` and `ID` in combination are a unique identifier for the item. Audios are converted to mono wav sampled at 16000 Hz.

The benchmark is created from a directory structure using the Datasets library, and subsequently restructured to use numerical labels, which can be mapped back to string labels.[6] This is required by subsequent steps. 15% of the data are separated into a validation split. In addition, a validation split is created for each label.[7] For instance, there is a validation split *validation_muchasidid0100*, with 250 tokens of *much+as+i+did* with expected subject prominence. These splits are used for obtaining classification accuracy for individual labels.

In order to bootstrap the format for the benchmark as it is being constructed, and possibly limit overfitting during training by increasing segmental diversity, nine words from the Superb KS benchmark are included in the benchmark (*down*, *left*, *no*, *off*, *on*, *right*, *stop*, *up*, and *yes*)[13]. We expect to remove these as the number of minimal pairs in the benchmark grows.

[6]See `prosodic_minimal_hub.py` in the code repository.

[7]See `prosodic_minimal_validations.py` in the code repository.

4. Machine learning classification

A standard paradigm for audio classification uses a dataset structured like the one from the previous section. Different vocabularies of labels may be involved, for instance environmental sounds (dog, rain, helicopter, etc.) in the ESC-50 dataset [14]; urban sounds (air conditioner, drilling, gunshot etc.) in the UrbanSound8K dataset [15]; isolated speech commands (up, down, on, off etc.) in Superb KS [13]; emotions (anger, happiness, sadness, etc.) in the IEMOCAP dataset [16]. Typically a base model is tuned on the training portion of a benchmark, and evaluated for accuracy in selecting the target labels for audios in a test set. Traditional sound representations such as raw waveform, spectrogram, or MFCC may be used, or alternatively representations learned with self-supervised neural models, such as wav2vec2 feature encoder outputs [17], HuBERT/WavLM representations [18, 19], or audio spectrogram transformer learned features [20]. We used wav2vec2 representations, a base model wav2vec2-base and an off-the-shelf procedure for training using the `run_audio_classification.py` program from the Github repository for [21]. Tuning an audio-classification model from wav2vec2-base works by iteratively adjusting the model's parameters so that the predicted labels match the true labels in the training data. Each audio clip is passed through the feature extractor and transformer layers to produce a vector representation, and a classifier head maps this to logits over the label set. A loss function measures the discrepancy between the predicted probabilities and the ground-truth labels. During gra-

label	train	test	accuracy
object10	1.4k	227	0.9885*
object01	1.18k	222	0.9671*
present10	898	161	0.9934*
present01	611	103	0.9661*
muchasidid0010	1.5k	255	0.96*
muchasidid0000	993	165	0.9818*
somepeople10	325	62	0.9677*
somepeople01	780	122	0.9508*

Figure 3: *Classification accuracies and sizes of train and test sets for four prosodic minimal pairs. Accuracies are likely overestimated (indicated by the asterisk), because there is likely to be some overlap between train and test sets, caused by the same audio occurring in different videos. This will increase accuracy, because the model can memorize a training example.*

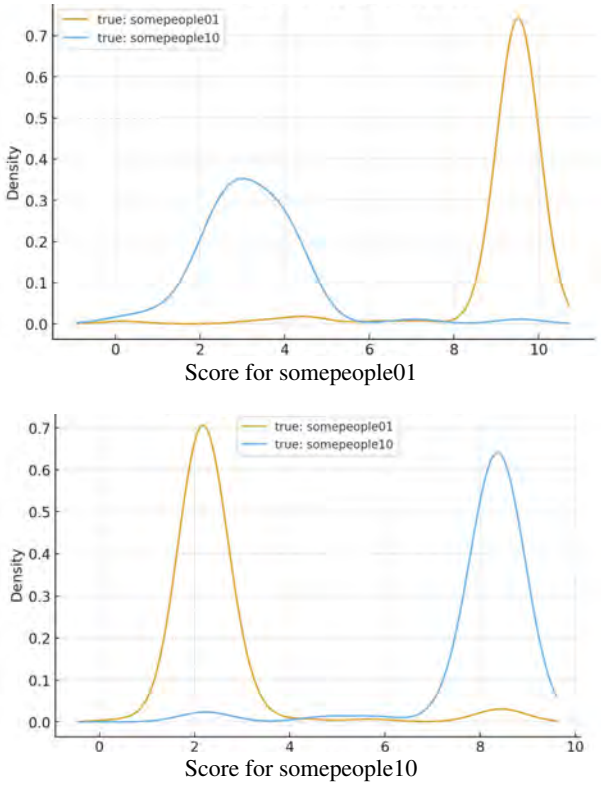

Figure 4: *Smoothed density plots indicating that the tuned model separates the two prosodies for* some+people *well. The plot on top graphs densities of scores for true instances the label* somepeople01 *(default prosody, in orange) and true instances the label* somepeople10 *(determiner focus, in blue) as realiations of the label* somepeople01. *Tokens that are more expected in the model of* somepeople01 *have higher scores.*

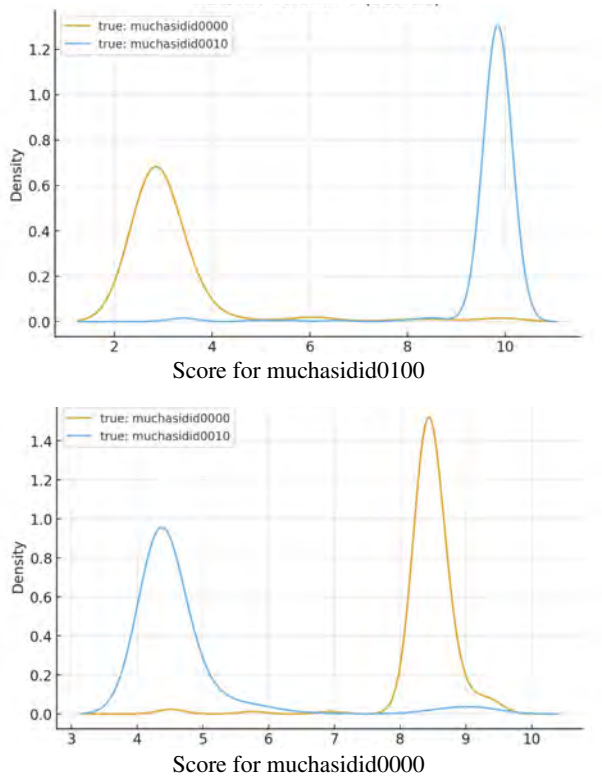

Figure 5: *Density plots of model scores for* muchasidid0100 *(subject focus) and* muchasidid0000 *(following focus).*

dient descent, the gradients of this loss with respect to all model parameters are computed via backpropagation, and an optimizer updates the weights in the direction that reduces the loss.

Since the current version of the benchmark includes no separate test data, training was simply run for ten epochs (complete passes over the training data), and the final model was adopted. Then the validation splits were used for testing. Separate tests were run for each prosodic reading of each minimal pair. In principle, it would be desirable to limit choices to the two labels for the pair, for instance *object01* and *object10* when classifying an item with true label *object01*. This is feasible, but has not been coded yet. Instead, classification used the full set of potential labels, with the model selecting the label with the highest score. In practice this makes no difference in our experiments, because examination of logs shows that the two highest-ranked labels were always the elements of the minimal pair. Figure 3 gives classification accuracies. Figure 4 gives density plots of scores for the opposition between the focus prosody and the default prosody of *some+people*. Figure 5 has density plots for *much+as+i+did*. For both minimal pairs, the model separates the two prosodies well.

5. Discussion

The paper illustrated that it is possible to retrieve on the order of a thousand audios of a short word sequence using the index at `filmot.com` of Youtube automatic transcriptions. We described a prototype benchmark of prosodic minimal pairs. The chief work that remains to be done on it is to eliminate duplicates coming from the same audio being used in different videos. This is underway using audio fingerprinting [22], and corrected results will be presented at the conference. Further, the benchmark should be expanded to twelve pairs or more, to include separate validation and test splits, and to have at least 1000 tokens for each type, in order to support experimentation with different machine learning approaches. Preliminary results indicate that a generic audio classification model can be tuned to make prosodic distinctions. To the best of our knowledge, the work here is the first large scale linguistic study to exploit Filmot (`filmot.com`) as a primary data resource.

Applications of the benchmark in various areas are anticipated. Investigations in machine learning can attempt to improve classification accuracy. Investigations of the linguistic theory of focus can attempt to verify whether predictions accord with naturalistic data, referring to model-generated judgments about whether a word is focused or not. The database and tuned model can be used to evaluate whether text to speech systems are producing the right prosody, by comparing the label in the benchmark to the label assigned by the model to audio synthesized from text in the benchmark.

6. Acknowledgements

Moti, the developer of `filmot.com`, provided important assistance in accessing the website from software.

7. References

[1] J. Howell and M. Rooth, "Web harvest of minimal intonational pairs," in *Web as Corpus 5*, 2009.

[2] J. Howell, "Meaning and Prosody: On the Web, in the Lab and from the Theorist's Armchair," Ph.D. dissertation, Cornell, 2011.

[3] J. Howell, M. Rooth, and M. Wagner, "Acoustic classification of focus: On the web and in the lab," *Laboratory Phonology*, vol. 8, no. 1, 2017.

[4] A. Rawat, V. Rawat, N. Singh, N. Kuchhal, J. Barmola, and H. S. Negi, "An enhance version of youtube video downloader using python," in *2023 International Conference on Computer Science and Emerging Technologies (CSET)*. IEEE, 2023, pp. 1–6.

[5] jianfch, "Stable-ts: Stabilizing timestamps for openai's whisper," https://github.com/jianfch/stable-ts, 2023, version 2.1.2, MIT License.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, vol. 1. Hawaii, 2011, pp. 5–1.

[8] L. Horn, "On the semantic properties of logical operators in english," Ph.D. dissertation, UCLA, 1972.

[9] M. Rooth, "Alternative semantics," in *The Oxford handbook of information structure*, 2016.

[10] D. Büring, "(contrastive) topic," in *The Oxford handbook of information structure*, 2016.

[11] O. Jespersen, *A Modern English Grammar on Historical Principles. Part VI: Morphology*. Copenhagen: Ejnar Munksgaard, 1941, reprinted by Routledge, London, 2007.

[12] Q. Lhoest, A. V. del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. L. Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. M. Rush, and T. Wolf, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2021, pp. 175–184.

[13] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, W.-C. Lai, K. Lakhotia, A. T. Lin, J. Liu, J.-H. Shi, X. Chang, G.-T. Huang *et al.*, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.

[14] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[15] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[18] W.-N. Hsu, B. Bolte, Y.-S. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6533–6547.

[19] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Chen, Z. Zhao, Y. Qian, J. Li, F. Wei, and X. Yao, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2022)*, 2022, pp. 150–154.

[20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proceedings of Interspeech 2021*, 2021, pp. 571–575.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[22] J. Six, "Panako: A scalable acoustic fingerprinting system handling time-scale and pitch modifications," *Journal of Open Source Software*, vol. 7, no. 73, p. 4554, 2022.