Speech-based Age and Gender Prediction with Transformers

Felix Burkhardt[1], *Johannes Wagner*[1], *Hagen Wierstorf*[1], *Florian Eyben*[1], *Björn Schuller,*[1,2,3]

[1]audEERING GmbH, Germany,
[2]Chair EIHW, University of Augsburg, Germany,
[3]GLAM, Imperial College London, UK

Abstract

We report on the curation of several publicly available datasets for age and gender prediction. Furthermore, we present experiments to predict age and gender with models based on a pre-trained wav2vec 2.0. Depending on the dataset, we achieve an MAE between 7.1 years and 10.8 years for age, and at least 91.1% ACC for gender (*female*, *male*, *child*). Compared to a modelling approach built on hand-crafted features, our proposed system shows an improvement of 9% UAR for age and 4% UAR for gender. To make our findings reproducible, we release the best performing model to the community as well as the sample lists of the data splits.

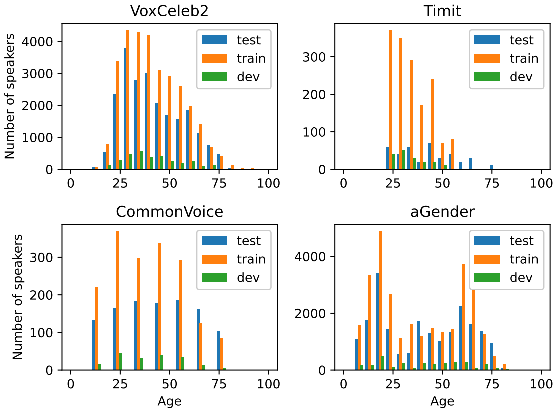

Figure 1: Distribution of speaker age (#samples) in the datasets for the three splits (CommonVoice age in mid-decades).

1 Introduction

The automatic detection of speaker age and gender has many use cases in human computer interaction, for example for dialogue adaption or market research. In contrast to subjective phenomena such as emotional arousal, the age of a person may be objectively determined, like for example body size, by an exact measurement. But, just like emotional arousal, age is only one of many factors that influence the acoustic speech signal [30], and typically not to be predicted to the year.

We curated several publicly available datasets with respect to age labels and used them to train several age models based on a wav2vec 2.0 architecture. We experiment on in- and cross-domain prediction, multi-head vs single head models and the number of transformer layers to be used. Finally we report the performance, using the 2010 paralinguistic challenge winner as a baseline.

Age and gender prediction based on machine learning as such has been investigated numerous times in the past decades. A problem in this regard is the lack of benchmark datasets that could be used to compare approaches. There are several publicly available age annotated datasets like the SpeechDat II corpus[1], CommonVoice [28], aGender [20], Timit [21], VoxCeleb2 [27], or the NIST test set [19], but the studies we found used only some of these, or not comparable train-development-test splits. In addition, the authors usually only report either regression or classification results, and use different metrics such as mean average error (MAE), accuracy (ACC), precision, recall, or unweighted average recall (UAR).

During the 2010 Interspeech Paralinguistic Challenge [22], age classification was one of the topics. Lingenfelser *et al.* [9] report on the aGender dataset by fusing the results of ensemble classifiers trained on subgroups of a larger feature set and get 42.4% UAR on four age groups, the baseline being 46.2% UAR. Katerenchuk [13] use a similar configuration with respect to classifiers and feature sets to fuse acoustic and metadata for child speech detection. Early studies are also based on the aGender dataset [4, 6], both using gaussian mixture models (GMM)/support vector machines (SVM) meta classifiers. In Metze *et al.* [3], the best system reaches an F1 value of .54 with a Linear Discriminative Analysis (LDA) on Hidden Markov Models (HMM)s modelling phonemes on the SpeechDat II corpus. A human evaluation on a subset of the data reaches .61 F1 value. Sadjadi *et al.* [18] describe joint gender and age estimation using ivectors and Support Vector Regression, reaching 4.7 years MAE on the NIST SRE 2010 telephony test set. Sánchez-Hevia *et al.* [15] and Tursunov *et al.* [31] both describe joint gender and age classification with a convolutional neural network (CNN) based on the CommonVoice dataset and report a recall of 76% on six mixed gender-age groups and 74% UAR on twelve mixed gender-age groups, respectively. Comparing linear and logistic regression on ivectors with CNNs using spectrograms, Hechmi *et al.* [16] report .98 F1 for gender classification and 9.44 years MAE for age regression. They use part of VoxCeleb2 as a dataset with age ground truth labels estimated mainly from a Wikipedia lookup, and these labels are the basis for the VoxCeleb2 data used in this paper, see Section 2.1 for details. Zazo *et al.* [19] propose a long short term memory (LSTM) recurrent network and report MAE of 6.58 years on the NIST test set. Gupta *et al.* [17] classified age using a wav2vec 2.0 model on the Timit dataset and report 5.54 years and 6.49 years MAE for male and female speakers, respectively. Kwasny and Hemmerling [33] reached a similar performance on Timit utilising a QuartzNet architecture, pre-trained on CommonVoice and VoxCeleb2, and fine-tuned on a joint age and sex prediction. With respect to gender prediction, authors usually refer to the biological sex. Levitan *et al.* [29] report on the aGender dataset three-class problem and reach 85% accuracy on the test set with a Random Forest classifier and MFCC features. Alnuaim *et al.* [32] use a pre-trained ResNet 50 and fine-tune it for gender on a balanced sub-set CommonVoice. They report a recall of .958 on two gender

[1] https://tinyurl.com/speechdatii

groups on the test set of CommonVoice and a similar recall for cross-corpus performance.

With the paper at hand we see the following contributions:

- We evaluate a novel system to estimate age and gender with fine tuned transformer models
- We present curated sample sets for train, development and test splits of publicly available data sets and make them available to the research community
- We compare a combined age and gender model to models specialized on a single task
- We present in-domain and cross-corpus results to examine the generalisability of the proposed system
- We investigate how many transformer layers are actually needed to properly model the tasks
- We release our best performing model to the public

2 Datasets

In order to test our approach on publicly available datasets, we considered the ones mentioned in the introductory Section 1. Most of them have drawbacks: The Timit and NIST datasets mainly contain young adults, Mozilla Common Voice is labelled with self-reported decades as age, and aGender is only available in 8 kHz telephone quality, VoxCeleb2 might contain samples from the same speakers, but recorded in different years. An overview on the age distributions per split can be seen in Figure 1 and the numbers of samples and speakers per dataset and split in Table 1. All datasets are available for non-commercial research and the file lists can be accessed in the GitHub repository that accompanies this paper.[2].

2.1 VoxCeleb2

Hechmi *et al.* [16] report on a dataset which is based on a self-collected table for VoxCeleb2 speakers. Because the authors (and the github repository) do not provide exact sample lists but only the speaker splits, we limited the number of samples per speaker to 20 samples. The original number of samples per speaker in the VoxCeleb2 dataset has a mean value of 220. We re-used the test set and split the train set randomly into 10% development speakers and 90% for training. The age distribution for train and test splits is shown in Figure 1. As said, we followed the splits used in Hechmi *et al.* [16] which have a rather large portion of test speakers. Although the age peak is still on the young side, it reflects the world age distribution better than, for example, Timit. As many samples are quite long (> 20 s), we used voice activity detection (VAD) to segment the samples.

Dataset	train	devel	test
VoxCeleb2	30300 (1515)	3120 (156)	22100 (1105)
CommonVoice	1729 (118)	186 (13)	1110 (79)
Timit	1570 (157)	170 (17)	380 (38)
aGender	29553 (324)	2974 (35)	20549 (239)

Table 1: Overview of the datasets: #samples and #speakers (in parenthesis).

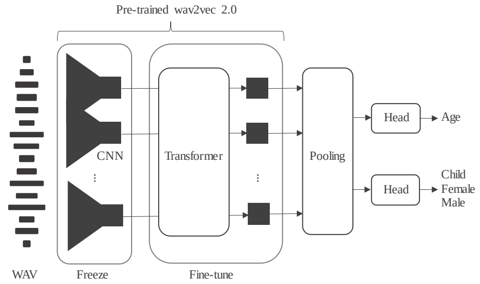

Figure 2: Proposed architecture built on wav2vec 2.0.

2.2 CommonVoice

The dataset is the result of a public data collection by the Mozilla foundation [28]. We used the *de-validated* collection as a basis for our datasets. Because the number of samples per speaker varies strongly (Mean: 55, STD: 296.6), we limited to 20 samples per speaker. Because the age distribution is quite imbalanced, we furthermore tried to age-balance the test and training splits by selecting at most 20 speakers per age decade and gender, and then chose at most 7 speakers per age-gender group as a test set and the others for training. As a development set, we randomly used 10% of the training speakers.

2.3 Timit

The well known Timit datasets [21] contains 16 kHz recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. Again, we tried to age-balance the test and training splits by first selecting at most 40 speakers per age decade and gender and then using 5 speakers per age-gender group as a test set, disregarding many of the speakers in their twenties. As a development set, we randomly used 10% of the training speakers.

2.4 aGender

The aGender dataset [20] has been collected via telephone especially with age classification as a focus. It has been used in the Interspeech 2010 Pralinguistic challenge[22]. The winning paper was Kockmann *et al.* [23] and they reached 53.86 UAR and 81.57 UAR for the age (4 classes) and gender (3 classes) classification for the development set respectively with a meta classification based on SVM and GMM. Because we wanted to be able to compare our results with the challenge, we used the official development set as our test set and took 10% of the training speakers as a development set (the test set of the aGender dataset is secret and not accessible). It is the only dataset that has been collected solely for the purpose of biological age prediction and therefore does contain a substantial amount of children and already a balanced age structure.

3 Architecture

Our proposed architecture is depicted in Figure 2. It is built on wav2vec 2.0 [25] with two custom heads to predict age and gender, respectively. We do not start training from scratch, but use the weights from a pre-trained model. In our experiments we rely on *wav2vec2-large-robust*[3], a model pre-trained on read speech from Libri-Light (60k hours) and CommonVoice (600 hours), but also

[2] https://github.com/audeering/w2v2-age-gender-how-to

[3] https://huggingface.co/facebook/wav2vec2-large-robust

noisy telephone speech from Fisher (2k hours) and Switchboard (300 hours) [26]. We could show that models fine-tuned on this variant are generally more robust against noise compared to models that have seen only clean speech during the pre-training [24].

As input to the heads we use the pooled hidden states (average pooling) of the last transformer layer. Each head consist of a fully connected layer of size 1024, a dropout layer, and a final projection layer. In case of age, we project to a single value predicting the *age* in range 0 to 1, where 1 corresponds to a hundred years. In case of gender, we project to three values expressing the confidence for being *child*, *female*, and *male*. During evaluation we decide in favor of the class with the highest value.

For fine-tuning on the downstream task, we use the ADAM optimiser with a fixed learning rate of 1e−4. Depending on the task we use two different loss functions: concordance correlation coefficient (CCC) loss for age, which we define as a regression problem; cross entropy (CE) loss for gender, which we define as a multi-class classification problem. For backpropagation we use the average of the two losses. We run for 5 epochs with a batch size of 64 and keep the checkpoint with best performance on the development set. As proposed in [**wagner2022dawn**] we freeze the CNN layers but fine-tune the transformer ones. When using the term fine-tuning, we will henceforth refer to this partial fine-tuning. These models are trained using a single random seed, for which the performance is reported.

In total, the model has 317.5M parameters. On a three second long input it performs 53.8G MAC operations, which took 34.2 ± 5.2 ms when measured on a NVIDIA RTX A4000 (100 repetitions). In 4.3 we will discuss how these numbers can be reduced. As learning framework we use PyTorch[4] and rely heavily on the transformer library by HuggingFace [34].

4 Experiments

We will now report results on predicting age and gender of a speaker from her or his voice. We treat gender detection as a classification task and report results in terms of accuracy (ACC) or unweighted average recall (UAR) for the three classes child, female, and male. In case of age, which we model as a regression problem, we report concordance correlation coefficient (CCC), and, for the aGender age groups: ACC or UAR.

4.1 Single vs combined model

We compare the performance of a *combined* model trained on both tasks simultaneously to that of models trained on a *single* task, i. e., either age or gender. For the latter, we use the same architecture described in Section 3 but remove the other head.

Results are summarised in Figure 3 and it is not difficult to recognise that the single and combined model perform almost identical. We can conclude that, although the combined model does not benefit from the information of the other channel, it is well able to learn both tasks at once. Since this (almost) halves the resources needed to run two separate models, a combined model should be preferred in a multi-task setup. Throughout the remaining of the paper, we will report results for the combined architecture without explicit mention.

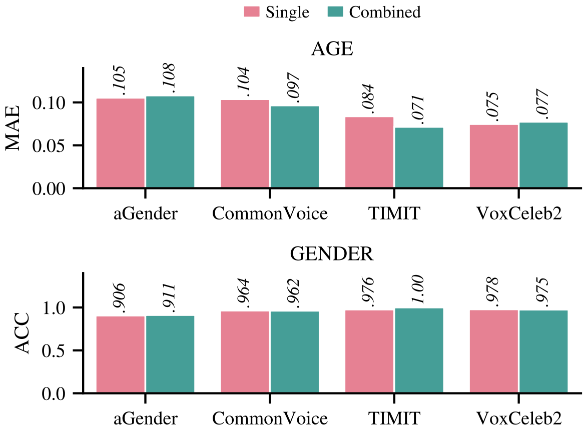

Figure 3: Performance of a model trained on a single task, i. e., either age (MAE in years/100) or gender, and a combined model simultaneously trained on both entities. We see that accuracy stays more or less the same.

4.2 Cross-corpus evaluation

In cross-corpus evaluation, a model is presented with data from an unknown source. To simulate such a situation, we train a model on a single dataset. For our experiment, we choose aGender as it is the only dataset that contains a considerable amount of child speech. In Figure 4, we report the performance of this model as *cross-corpus* on the remaining datasets, namely CommonVoice, Timit, and VoxCeleb2. For comparison, we also include *in-domain* results by the model trained on all datasets.

Removing in-domain data from the training generally leads to a performance drop on all datasets. On CommonVoice, the effect is yet small: plus one year for age and minus one percentage point for gender. Whereas on Timit and VoxCeleb2 it is quite significant: MAE increases by four to five years for age and ACC decreases by more than ten percentage points for gender. In the lower part of Figure 4, we visualise gender and age predictions aggregated over the three datasets. The confusion matrices reveal that the cross-domain model has problems in predicting females, while the distribution plots show that age is generally underestimated.

We conclude more and diverse data is needed to build a robust age and gender model, especially when using an upsampled 8 kHz dataset as training.

4.3 Varying the number of layers

In the experiments reported so far, we have used all 24 transformer layers. Dropping some of top layers reduces the footprint of a model. However, too few layers may degrade the ability of the model to properly learn a task. To investigate the effect of reducing transformer layers, we run experiments with a varying number of layers. As we see in Figure 5, results generally improve with more layers, though the effect is smaller for gender than for age. In fact, a single transformer layer seems sufficient to adequately model gender. For age, we can observe a considerable performance drop for less than six layers, while with more than six layers, there is only a marginal increase.

We conclude that using six transformer layers provides a good trade-off between accuracy and speed. This reduces the number of parameters by a factor of 3.5 to 90.8M and inference time by a factor of to 3 to 12.4 ± 3.3ms.

[4] https://pytorch.org/

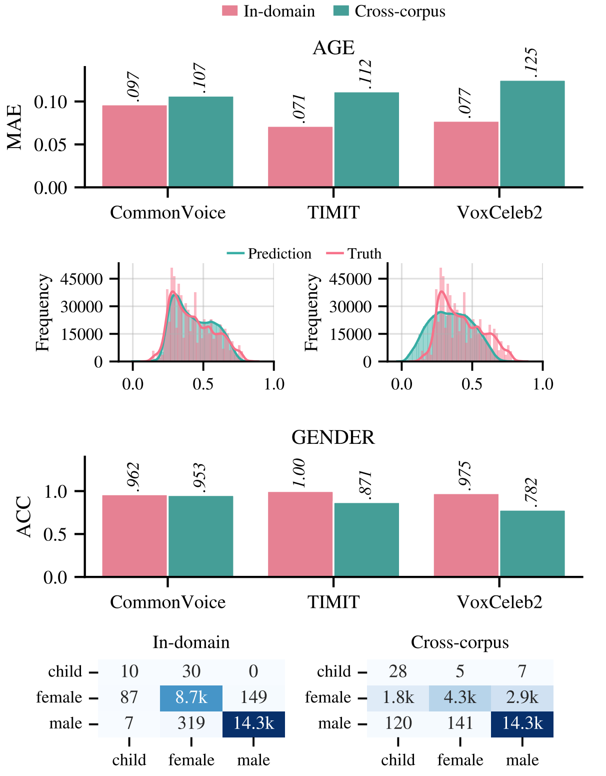

Figure 4: In the cross-corpus condition, the model has not seen data from the dataset it is evaluated on. Results are compared with a model trained in an in-domain fashion. The cross-domain model performs poorly in detecting females (cf. confusion matrices) and predicts speakers younger than they are (cf. distribution plots) (age: MAE in years/100).

4.4 Comparison classical modelling approach

Finally, we compare performance to a classic modelling approach based on hand-crafted features. As a baseline, we choose the winner system of the 2010 Interspeech Paralinguistic Challenge [22]. It implements a combination of Gaussian Mixture Models (GMM) and Support Vector Machines (SVM), followed by linear Gaussian backends and logistic regression-based fusion, which uses as input a large feature set of acoustic, prosodic, and voice quality features. For more information see Kockmann *et al.* [23].

In Table 2, we compare the performance of the baseline system with our best performing model (24 layers) when trained either only on aGender or on all datasets introduced in Section 2. In case of age, we map the continuous predictions of our model to the four classes child, youth, adulate, and senior as proposed by the challenge organisers. In addition, we include results for the combined age/gender task with seven classes used in the challenge [22]. Depending on the task, UAR and ACC of the baseline is improved by 4–10 percentage points. Using all datasets during training provides an additional, yet small boost.

We can conclude that deep learning improves the accuracy compared to a classic modelling based on manual feature engineering.

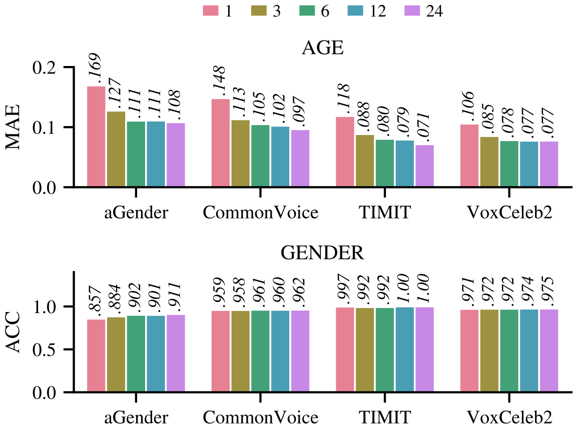

Figure 5: Results with different number of transformer layers. For gender, a single transformer layer seems sufficient, whereas for age, six layers provide a good trade-off between speed and accuracy (age: MAE in years/100).

System	Age 4-class	Gender 3-class	Combined 7-class	Training
		Development		
baseline [23]	.56 / .55	.82 / .87	.54 / .54	aGender
wav2vec 2.0	.60 / .60	.84 / **.92**	.58 / .60	aGender
wav2vec 2.0	**.61** / **.61**	**.86** / .91	**.59** / **.61**	All
		Test		
baseline [23]	.52 / .51	.83 / .86	*N/A*	aGender
wav2vec 2.0	.60 / **.57**	.86 / **.90**	**.57** / **.56**	aGender
wav2vec 2.0	**.61** / **.57**	**.87** / .88	**.57** / **.56**	All

Table 2: Comparison with baseline system based on hand-crafted features. Results are reported in terms of UAR / ACC on the development set (upper part) and test set (lower part). In case of age, we map the predictions of our model to the four classes child, youth, adult, and senior. In the combined task, age and gender are jointly represented by seven classes.

4.5 Age/gender prediction of emotional data

When testing the published model on the Berlin emotional database [35], the MAE for all samples age prediction is 8.35 and the UAR for binary gender prediction 96.04. When only the neutral samples are used, the MAE drops to 5.94 and the UAR to 100. Clearly, the acted emotional expression jeopardizes the quality of our model, as it has been trained on non-emotional data.

5 Conclusions and Outlook

We performed experiments on age and gender prediction based on four datasets and a fine tuned transformer architecture.A model trained on all data sets, together with the test, train, and develpment splits, has been made public and can be used as a baseline for other authors. We will continue to investigate age detection by using other model architectures and perhaps combining them with expert features. Especially speech data from children is sparse and we will look for such data or try to synthesise data with generative models.

6 Acknowledgements

This research has been partly funded by the European EASIER (Intelligent Automatic Sign Language Translation) project (Grant Agreement number: 101016982) as well as the European MARVEL (Multimodal Extreme Scale Data Analytics for Smart Cities Environments) project (Grant Agreement ID: 975337).

7 References

[1] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE — the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462, ISBN: 978-1-60558-933-6.

[2] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg J.and Burgoon, A. Baird, A. Elkins, Y. Zhang1, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, 2016.

[3] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. Bauer, and B. Littel, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 4, 2007.

[4] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, 2008.

[5] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[6] M. Feld, F. Burkhardt, and C. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[7] M. Brückl, *Altersbedingte Veränderungen der Stimme und Sprechweise von Frauen* (Mündliche Kommunikation). Berlin: Logos Verlag, 2011, vol. 7.

[8] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[9] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, "Age and gender classification from speech using decision level fusion and ensemble based techniques," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 2002.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2016.

[12] M. Brückl and F. Heuer, *Irrna: Coefficients of interrater reliability – generalized for randomly incomplete datasets*, R package version 0.1.4, 2018. [Online]. Available: https://CRAN.R-project.org/package=irrNA.

[13] D. Katerenchuk, "Age group classification with speech and metadata multimodality fusion," in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2017.

[14] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions.," in *INTERSPEECH*, 2019, pp. 3935–3939.

[15] H. A. Sánchez-Hevia, R. Gil-Pita, ·. M. Utrilla-Manso, M. Rosa-Zurera, and M. Utrilla-Manso, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, 2022.

[16] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 687–693.

[17] T. Gupta, D. T. Truong, T. T. Anh, and C. E. Siong, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September, pp. 1978–1982, Mar. 2022.

[18] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 5040–5044, May 2016.

[19] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on lstm recurrent neural networks," *IEEE Access*, vol. 6, pp. 22 524–22 530, Mar. 2018.

[20] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *Darpa timit acoustic phonetic continuous speech corpus cdrom*, 1993.

[22] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pp. 2794–2797, 2010.

[23] M. Kockmann, L. Burget, and J. Černocký, "Brno university of technology system for interspeech 2010 paralinguistic challenge," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, vol. 2010, Makuhari, Chiba, JP: International Speech Communication Association, 2010, pp. 2822–2825.

[24] 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*, pp. 1–13, 2023.

[25] 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 (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 12 449–12 460.

[26] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[27] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101 027, Mar. 2020.

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 4218–4222, 2020.

[29] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," *Proceedings of the International Conference on Speech Prosody*, vol. 2016-January, pp. 84–88, 2016.

[30] S. Schötz, "Acoustic analysis of adult speaker age," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4343 LNAI, pp. 88–107, 2007.

[31] A. Tursunov, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, 2021.

[32] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Speaker gender recognition based on deep neural networks and resnet50," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–13, 2022.

[33] D. Kwasny and D. Hemmerling, "Joint gender and age estimation based on speech signals using x-vectors and transfer learning," *arXiv preprint arXiv:2012.01551*, 2020.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[35] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *9th European Conference on Speech Communication and Technology*, vol. 5, Sep. 2005, pp. 1517–1520. DOI: 10.21437/Interspeech.2005-446.