Proceedings Paper:
Ellis, S., Goetze, S. orcid.org/0000-0003-1044-7343 and Christensen, H. (2023) Moving towards non-binary gender Identification via analysis of system errors in binary gender classification. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023), 04-10 Jun 2023, Rhodes Island, Greece. Institute of Electrical and Electronics Engineers . ISBN 9781728163284

https://doi.org/10.1109/ICASSP49357.2023.10095997

MOVING TOWARDS NON-BINARY GENDER IDENTIFICATION VIA ANALYSIS OF SYSTEM ERRORS IN BINARY GENDER CLASSIFICATION

Sebastian Ellis, Stefan Goetze and Heidi Christensen

Speech and Hearing Group, Dept. of Computer Science, Sheffield, UK
{scgellis1, s.goetze, heidi.christensen}@sheffield.ac.uk

ABSTRACT

This paper aims to analyse human perceptions of gender in speech signals, focusing on signals that are misclassified by methods for binary gender classification, looking at the features of speech signals that are more likely to be misclassified, or classified as either nonbinary or unclassifiable. The paper also analyses how human subjects perform in classifying such speech signals to gain insight into differences between machine and human performance levels. It is shown that gender classification systems and human ratings lack inter-annotator agreement, as do human ratings considered individually. There is also discussion of the suitability of continuing to use a binary system for gender in the field. This work fits into a larger body of research ongoing in the area of speech technology for transgender voice therapy.

Index Terms— Binary Gender Classification, Human Evaluation, Transgender Voice

1. INTRODUCTION

Gender is one of the most commonly available metadata in public speech datasets. Gender itself is also the focus of several speech technology tasks such as estimating the binary gender of a speaker [1–3], and anonymising a speaker through gender deidentification [4–6]. State-of-the-art systems have achieved scores very close to 100% recognition rate [1]. However, open challenges remain.

An increasing amount of scientific literature highlights the issues with the idea of a *gender binary* (i.e., classifying people into two opposite genders) rather than a spectrum [7]. Despite this, binary gender classification systems are still widely used and have been for some time, with examples in the speech domain including [8–11] and even further afield in domains such as facial recognition [12]. Reasons for using these systems include call-centre routing, demographic tracking, forensics, and personalised marketing. These systems can also be used in applications aimed at gender / speech anonymisation [5].

In general, the idea of a gender binary is becoming less and less relevant in general society. Up to 5% of Generation-Z youths identify as non-binary [13, 14], meaning that they feel that their gender identity does not lie within typical conventional binary options. Furthermore, an increasing number of people are identifying as transgender [15]; more than 1.6 million Americans and up to 3% of New York youths. In addition to issues with a gender binary, there is also research to show that acoustic variations in speech with respect to gender are at least partially due to both the speaker's language and social constructs [16]. This would mean that both perceived and acoustic differences in speech signals from gender may shift over time as social views change, for example, possibly narrowing the gap in pitch differences in speech between genders [17].

Another area where voice and gender are in increasing focus is that of transgender voice therapy, which is used to alleviate voice-based dysphoria in individuals whose gender identity does not align with the gender they were assigned at birth. Therapy is the preferred option to other interventions such as surgery, as there is evidence that therapy alone can achieve results that most are happy with much lower risks - those who have opted to have surgery are also additionally offered therapy. However, voice therapy and other similar therapies are underfunded, leading to a number of issues relating to shorter sessions and a lack of support outside of sessions. Technical interventions in the form of, e.g., speech therapy tools could be used to provide augmented and extended care for these groups of potentially very vulnerable people.

In order to successfully develop technical interventions in this field (e.g., training systems for voice therapy) and integrate speech technology into therapy practices, it is necessary to better understand what aspects of speech signals make humans perceive gender - such as, perception of fricatives like /s/ and /ʃ/ [18]. From this it may be possible to further understand how computers and humans perceive gender in speech signals, and understand how current gender identifications work (and fail), as well as establish the features of speech signals which are more likely to be misclassified, or classified as either nonbinary or unclassifiable.

Attempting to solve or even fully explain the issues relating to the use of gender recognition systems is beyond the scope of this paper. Instead, the aim of this paper is to primarily analyse the binary gender identification system proposed in [1], to analyse the types of speech signals that it struggles to correctly identify and to investigate the characteristics of these misidentified speech signals as, for example, spectograms, looking for any common themes. In addition, the paper investigates how humans perform when asked to classify such misclassified speech signals in an attempt to compare the system with human performance levels. Finally, the paper will discuss thoughts on the suitability of continuing to use a binary system for gender in the field, looking at potential alternatives informed by both the collected human data and other sources.

The remainder of this paper is structured as follows: Section 2 introduces the system used to generate misclassified voice speech signals, then Section 3 introduces the datasets used and the listening experiment. Results are presented in Section 4 and the paper then concludes in Section 5.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

2. METHODOLOGY

The d-vector based gender estimation model proposed in [1] is taken as a baseline. It was selected as a state-of-the-art system with the aim to analyse its misclassified utterances. A high-level overview of the system is illustrated in Fig. 1.

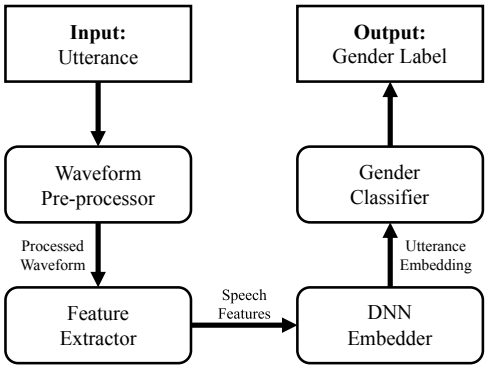

Fig. 1: High-level representation of the d-vector based baseline gender classification system.

The system uses a d-vector architecture, first introduced in [19], to embed the utterance through the use of a multi layer long short-term memory (LSTM) recurrent neural network (RNN). An input speech utterance is passed through a waveform preprocessor to standardise the files across the three datasets in terms of file format and peak loudness, before being passed to a feature extractor pretrained on the Voxceleb1 [20] and Librispeech [21] datasets. These features are passed to a DNN embedder, which creates a 256-element vector to pass to the gender classifier. The embedder's architecture is composed of three LTSM layers connected in series, each with a layer size of 256, which are then connected to a dense and ReLu layer.

As shown in Fig. 1, there is an additional neural network-based binary gender classifier on top of the d-vector embedder. Following the procedure in [1], this network was pre-trained on the VoxCeleb1 dataset [20] and then subsequently fine-tuned on both CommonVoice [22] and DARPA-TIMIT [23]. The architecture of the binary gender classifier in Fig.1 is presented in Table 1.

Table 1: A summary of the system's binary gender classifier.

Layer	Input Size	Output Size
Dense + ReLu + BatchNorm	256	256
Dense + Logits	256	1

3. EXPERIMENTAL SETUP

3.1. Baseline Gender Classification System

The accuracy of the re-implemented system depicted in Fig.1 is 99.29%, which is slightly lower (0.31%) but approximately in line with the results achieved in [1]. Although unspecified in [1], the implemented baseline system uses a binary cross-entropy loss function. The model is trained and fine-tuned on multiple datasets as in [1] as follows:

VoxCeleb1 [20] is a large-scale text-independent speaker identification dataset stored in the ".wav" format, collected from pre-existing footage available on YouTube. It features over 7000 speakers, 1 million utterances and 2000 hours of data. The available metadata for the gender of the speaker was originally created via the use of a facial-recognition based convolutional neural network (CNN), which other papers have previously suggested may have lead to a large number of inaccurate gender labels [24].

CommonVoice [22] is a large-scale multilingual collection of transcribed speech designed primarily for automatic speech recognition (ASR) purposes stored in ".mp3" format, featuring more than 50,000 individuals and over 2,500 hours of audio. Demographic metadata is optionally self-provided by users recording their own voice. The dataset itself is also split by language; here only English speech is used (as in the original implementation [19]).

DARPA-TIMIT [23], or, the Defense Advanced Research Projects Agency (DARPA) Texas Instruments / Massachusetts Institute of Technology (TIMIT) corpus of read speech stored in the NIST-SPHERE format, was designed to provide speech data for early automatic speech recognition systems. TIMIT contains 630 speakers across eight dialects, each speaking 10 phonetically rich sentences. Metadata was collected directly from participants at the time of recording.

3.2. Gender Perception Experiment

Following both the training and the fine-tuning of the model, a total of 408 speech signals were misclassified in the dataset, which were combined with 100 correctly classified "control" speech signals, to provide an anchor for comparison with other speech signals.

A listening experiment was conducted with self-reported normal hearing participants who were presented with a single speech signal at a time and were asked to rate it on a continuous sliding scale. One end of the scale was labelled "Male", and the other "Female", with the slider starting in the center for each clip. Below this, there was a second slider representing the user's confidence in their gender rating of the speech signal, ranging from 0% to 100%. To speed up the participant's workflow, the slider would automatically move depending on the distance from the centre the participant had set the first slider; however, it could be adjusted individually as well - this was made clear to the participants. Participants were asked to rate at least 100 speech signals if possible, but were otherwise given no instruction on what a typical *male* or typical *female* sounded like, in order to keep results true to their own personal perception.

Overall, the 18 participants provided 2411 responses (Mean: 134, SD: 108), a maximum of 508 and a minimum of 20. The participants had a mean age of 28.5 (SD: 7.5), with 5 listing their gender as "Female" and 13 listing their gender as "Male".

4. RESULTS

As previously stated, the results of the baseline system are comparable although slightly lower than the original system presented in [1]. Since analysis of the listening test conducted by the human participants are the paper's main focus, Table 2 summarises a number of metrics analysing the listening test data.

In Table 2, BCE represents the binary cross-entropy loss among only human results and is intended as an intuitive comparison between the computer's performance. Similarly, two measures of accuracy are reported: $\mathrm{Accuracy}$ (Representing the standard accuracy measure) and $\mathrm{Accuracy_c}$ (Representing accuracy weighted by confidence), which allows us to establish a level of agreement between

Table 2: A summary of listening test results.

Metric	Experiment	Control	Combined
BCE	0.526	0.068	0.595
Accuracy	0.945	0.979	0.952
Accuracy_c	0.954	0.984	0.960
$\kappa_{\text{(human-only)}}$	0.096	0.440	0.179
$\kappa_{\text{(both)}}$	-0.056	0.497	0.039

humans and the base labels, in addition to approximating human error. Finally, Fleiss' Kappa, denoted by κ in table 2, is presented to highlight inter-annotator agreement between solely humans, and both the humans and baseline system.

For BCE, the loss was measured using the rating that each participant gave compared to the labels provided with the dataset:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (1)$$

for each corpus label y_i of the (mis)classified samples, giving rise to the probability that the label will be chosen $p(y_i)$ and a dataset of size N. Here, $N = 508$, the number of tracks in the dataset. The somewhat high BCE score in Table 2 may indicate that human participants were less confident in their labelling.

Accuracy represents the standard accuracy measure widely used,

$$\text{Accuracy} = \frac{\sum_{i=1}^{M}(\text{g}(p(y_i))}{M} \quad (2)$$

with $\text{g}(p(y_i))$ representing the number of correct results, e.g.

$$\text{g}(p(y_i)) = 1 \text{ if } p(y_i) > 0.5 \text{ else } \text{g}(p(y_i)) = 0 \quad (3)$$

and M represents the total number of ratings available (Here, $M = 2411$). Similarly to the above, $Accuracy_c$ represents a weighted accuracy score using confidence scores,

$$\text{Accuracy}_\text{c} = \frac{\sum_{i=1}^{M}(\text{g}(p(y_i)) \cdot c_i)}{\sum_{i=1}^{M} c_i} \quad (4)$$

for confidence score c_i, and provided rating $p(y_i)$ from each participant i, with

Accuracy is included as an intuitive rating of human performance over the data, and $Accuracy_c$ provides similar information but quantified by how confident the human participants were in their ratings. Interestingly, a perfect 100% accuracy score was not achieved. One reason for this could be human error; However, especially given the fact that VoxCeleb was initially classified by an automatic gender classification system itself, it may indicate that some labels on the data are incorrect. No participant achieved 100% accuracy (Mean: 0.944, SD: 0.042, Max: 0.987, Min: 0.820) or accuracy_c (Mean: 0.954, SD: 0.038, Max: 0.985, Min: 0.832). When taking into account the control samples alone however, 9 participants achieved a perfect accuracy and accuracy_c score. Analysing tracks which were misclassified, a mean disagreement with the label of 0.378 (SD: 0.298) can be observed, and a total of 10 tracks had 100% disagreement with the source label.

κ in Table 2 represents the Fleiss' Kappa coefficient which evaluates the inter-rater agreement,

$$\kappa = \frac{\hat{P} - \hat{P}_e}{1 - \hat{P}_e} \quad (5)$$

where $1 - \hat{P}_e$ is the degree of achievement available above simple random chance and $\hat{P} - \hat{P}_e$ is the degree of achievement actually achieved above random chance. $\kappa(human\text{-}only)$ is the metric evaluated excluding the computer's results, and $\kappa(both)$ is the metric evaluated including the computer's results. The "human-only" results are around 0.2, indicating a fair level of agreement, whereas the "both" results are close to 0 (poor agreement. As expected, the human-only agreement is higher than the human-computer agreement for the combined case, however the gap between the two scores is lower than expected.

4.1. Confidence Scores

As mentioned in Section 3.2, participants were asked to provide confidence scores along with their gender rating for each speech signal. Fig. 2 and later Fig. 3 show the distribution of these confidence scores.

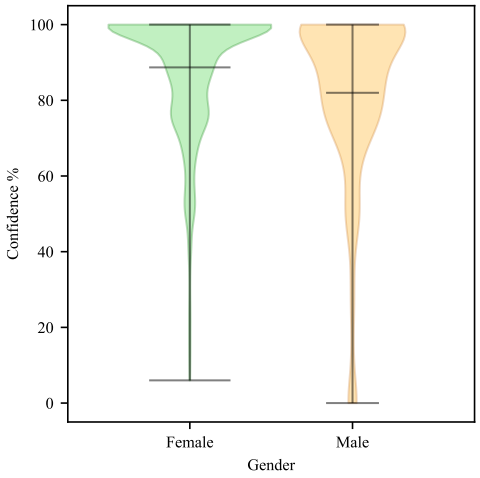

Fig. 2: A violin chart showing the frequency of confidence scores, with lines representing their range and means for each binary gender.

Fig. 2 shows us that the majority of confidence ratings were above 60%, and that participants were more often very confident when classifying female-labelled voices. Of particular note in Fig. 3 are interesting results as confidence scores increase. Two immediately obvious paths of a higher confidence relating to an explicitly male or explicitly female voice, but looking at participant gender ratings slightly to the right of a neutral rating, for corresponding confidence scores $\approx [85 : 95]$, there is a set of high-confidence ratings which are trending towards a high confidence gender neutral rating. This, combined with the ratings which are at 100% confidence for a gender rating of 0.5 (i.e. exactly between binary *male* and *female*) indicates that there is support for the theory that people may classify speech signals as nonbinary - that is, not as a particular binary gender.

Two distinct straight-line paths originating from confidence 0% to the opposite two corners of Fig. 3 can be observed. These represent participants not moving the confidence slider from its automatic position. Therefore, these data points are considered to be of less

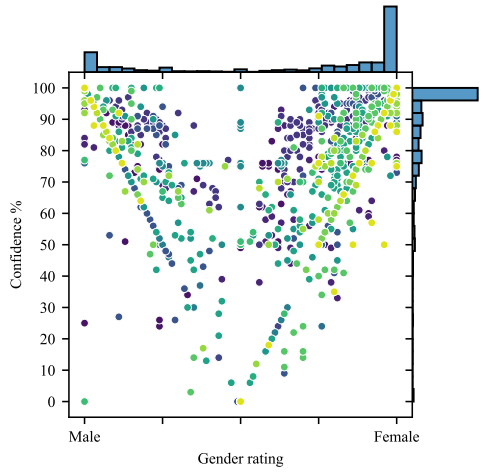

Fig. 3: A scatter plot showing the confidence scores for each continuous gender rating, coloured per participant. Histograms are depicted at each axis to visualise the data distribution.

interest. For repeating this experiment it may be interesting to investigate the effects of not pre-moving the confidence slider.

Overall, it can be noted that there are high confidences ratings for gender ratings that are not either 100% male or 100% female, and bearing the above in mind, there is evidence to support that humans are not thinking strictly in terms of binary gender. This finding, alongside emerging acceptance and narratives surrounding transgender people, may support the idea that implementing binary gender classification systems is increasingly irrelevant and possibly even unethical in today's world.

4.2. Result Regression

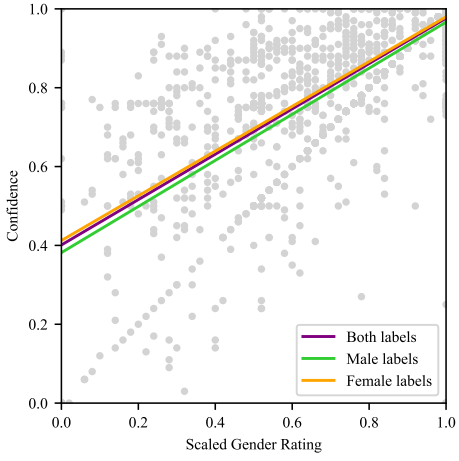

Fig. 4: A linear regression and scatter plot, fit to participant ratings

Fig. 4 shows the results of a simple linear regression model trained on the participant's gender ratings and confidence scores. The gender ratings, previously on a continuous scale from 0 (male) to 1 (female), were rescaled by the formula

$$\text{NewScore} = 2 \cdot \text{abs}(\text{OldScore} - 0.5) \quad (6)$$

with $\text{abs}(x)$ representing the absolute value of x. The data shows a clear correlation between a more gendered rating (*male* or *female*) and a higher confidence. However, of note is that this regression implies that even for non-binary gender ratings (e.g. close to the centre of *male* and *female*), that confidence scores remain high at approximately 0.5.

4.3. Example Spectograms

Fig. 5 shows four sample spectograms randomly chosen from the data used as part of this experiment, with accompanying labels. There are two correctly classified spectograms (ending "_C"), one for both "male" and "female", respectively (beginning "M" and "F" respectively), and two incorrectly classified spectograms (ending "_I"), again one for each gender in the dataset. In addition, each spectogram also lists the time-averaged pitch frequency F0.

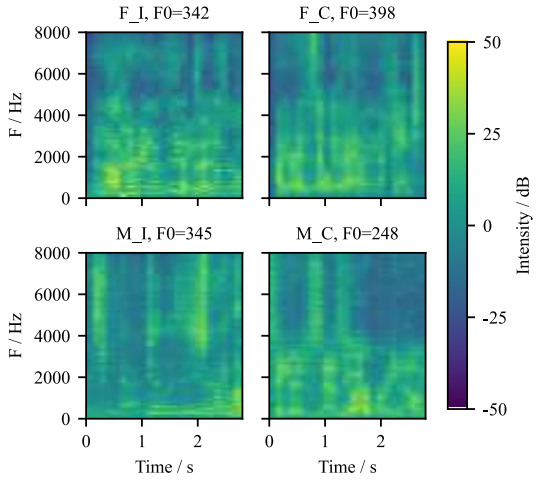

Fig. 5: Spectogram examples showing correctly classified and misclassified speech signals, along with their average F0 values.

A visual examination of the spectograms presents some interesting observations. Firstly, looking at the misclassified female speech signals in F_I, it can be seen that there are some similarities with the correctly classified male speech signals in M_C. Namely, both feature prominent high-energy bands in lower frequencies, and tend to retain energy into the higher frequencies. Meanwhile, comparing the misclassified male speech signals in M_I with the correctly classified female speech signals in F_C, the opposite of the above is true, lower energy in very high frequencies. Additionally, misclassified signals have similar F0s, close to 340Hz.

5. CONCLUSION

This paper analysed speech signals misclassified by a state-of-the-art binary gender classification system and analysed respective human assessment. A lack of agreement, between both human raters and classifiers, and between differing human raters was observed. This suggests that determining gender from voice is not something that humans are capable of doing with perfect agreement, as least based on solely listening to speech signals - a statement which is further supported by ratings indicating with high confidence that a speech signal was neither male nor female. Future work will look at what this means for automatic voice therapy tools helping people with dysphoria.

6. REFERENCES

[1] Damian Kwasny and Daria Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors*, vol. 21, no. 14, pp. 4785, 2021.

[2] Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Sikonja, "Multi-aspect multilingual and cross-lingual parliamentary speech analysis," *arXiv preprint arXiv:2207.01054*, arXiv, 2022.

[3] Marcos Faundez-Zanuy, Enric Sesa-Nogueras, and Stefano Marinozzi, "Speaker identification experiments under gender de-identification," in *2015 International Carnahan Conference on Security Technology (ICCST)*. Sept. 2015, IEEE.

[4] Stefano Marinozzi and Marcos Faundez Zanuy, "Digital speech algorithms for speaker de-identification," in *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)*. Nov. 2014, IEEE.

[5] Dimitrios Stoidis and Andrea Cavallaro, *Generating gender-ambiguous voices for privacy-preserving speech recognition*, arXiv, 2022.

[6] Brij Mohan Lal Srivastava, Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, and Marc Tommasi, "Design choices for x-vector based speaker anonymization," arXiv, 2020.

[7] Claire Ainsworth, "Sex redefined," *Nature*, vol. 518, no. 7539, pp. 288–291, Feb. 2015.

[8] D.G. Childers, Ke Wu, K.S. Bae, and D.M. Hicks, "Automatic recognition of gender by voice," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE Computer Society, 1988, pp. 603–604.

[9] R. Vergin, A. Farhat, and D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 1996, vol. 2, pp. 1081–1084 vol.2.

[10] Kavita Chachadi and S. R. Nirmala, "Gender recognition from speech signal using 1-d cnn," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, Vinit Kumar Gunjan and Jacek M. Zurada, Eds., Singapore, 2022, pp. 349–360, Springer Nature Singapore.

[11] Kavita Chachadi and S.R. Nirmala, "Voice-based gender recognition using neural network," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, pp. 741–749. Springer, 2022.

[12] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu, "Human gender classification: a review.," *Int. J. Biom.*, vol. 8, no. 3/4, pp. 275–300, 2016.

[13] The Trevor Project, "Diversity of nonbinary youth: July research brief - the trevor project," *The Trevor Project*, The Trevor Project, July 2021.

[14] Jeffrey M. Jones, "Lgbt identification in u.s. ticks up to 7.1%," *Gallup.com*, Gallup, June 2022.

[15] Jody L Herman, Andrew R Flores, and Kathryn K O'Neill, "How many adults and youth identify as transgender in the united states?," *Williams Institute*, Williams Institute, June 2022.

[16] Erwan Pépiot, "Voice, speech and gender: Male-female acoustic differences and cross-language variation in english and french speakers," *Corela*, June 2012.

[17] Cecilia Pemberton, Paul McCormack, and Alison Russell, "Have women's voices lowered across time? a cross sectional study of australian women's voices," *Journal of Voice*, vol. 12, no. 2, pp. 208–213, 1998.

[18] Elizabeth A Strand, "Uncovering the role of gender stereotypes in speech perception," *Journal of language and social psychology*, Sage Publications Sage CA: Thousand Oaks, CA, vol. 18, no. 1, pp. 86–100, 1999.

[19] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *CoRR*, vol. abs/1912.06670, 2019.

[23] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proceedings of DARPA workshop on speech recognition*, 1986, pp. 93–99.

[24] Khaled Hechmi, Trung Ngo Trong, Ville Hautamäki, and Tomi Kinnunen, "Voxceleb enrichment for age and gender recognition," *CoRR*, vol. abs/2109.13510, 2021.