

TurkSV-Bench: Assessing Demographic Bias and Preprocessing Effects in Speaker Verification

Ahmet Berkay Kantarcı
Ostim Technical University
220172022@ostim.teknik.edu.tr

Abdulkadir Parlak
abdulkadirparlak3@gmail.com

Ahmet Karazor
TURKSAT
ahmet.karazor@turksat.com.tr

Abstract

In this study, TurkSV-Bench is introduced as a novel benchmark dataset to evaluate the zero-shot cross-lingual performance and demographic biases of Turkish speaker verification (SV) systems. Created via an automated pipeline from YouTube data, this dataset comprises 1,844 audio recordings from 461 speakers representing diverse age groups, genders, and regional dialects. Evaluation of existing models identifies ResNet34-LM as the most successful baseline architecture. Subgroup analyses demonstrate the models' generalization capabilities across different gender and age demographics, while revealing distinct performance variations among regional dialects. Comprehensive evaluations of various preprocessing techniques highlight the baseline architecture's structural robustness, identifying specific configurations that maintain or refine verification performance without requiring aggressive signal filtering.

1 Introduction

Speaker verification (SV) is a biometric task that verifies whether a given voice recording belongs to a claimed identity by comparing it with a previously recorded reference sample. Unlike speaker identification, SV does not aim to recognize a specific speaker, but rather determines whether two voice samples belong to the same person. Today, SV systems are widely used in applications such as telephone banking, call center identity verification, smart assistants, mobile devices, access control, and digital forensics.

With the development of deep learning methods, embedding-based architectures have become the standard, transforming speaker-specific discriminative features into low-dimensional representation vectors and achieving high success on benchmark datasets such as VoxCeleb (Nagrani et al., 2017) and CN-Celeb (Fan et al., 2020).

However, the vast majority of existing speaker verification models are trained on English or multilingual datasets, making their application to Turkish speech inherently a cross-lingual transfer problem. Turkish differs significantly from these training languages in terms of phonetic structure, agglutinative characteristics, speech rhythm, and rich regional dialect diversity. Despite this, comprehensive studies systematically examining the performance of state-of-the-art speaker verification models on Turkish remain quite limited, making the evaluation of existing models' generalizability to Turkish speech an important research problem.

Another significant shortcoming in Turkish speaker verification research is the lack of a comprehensive and reliable benchmark dataset, as existing datasets are limited in terms of age, gender, regional dialects, and demographic diversity, making it difficult to systematically evaluate model performance across different user groups.

In this study, we develop an end-to-end automated data collection and processing pipeline for Turkish speaker verification and use it to create a new benchmark dataset, named as TurkSV-Bench. The system automatically collects audio recordings from publicly available YouTube videos and playlists, performing voice activity detection, speaker diarization, removal of overlapping speech, quality control, and data management. The dataset accounts for varying age groups, genders, and regional dialect diversity across Turkey, yielding a traceable benchmark for analyzing both speaker verification performance and potential demographic biases.

Rather than proposing a new speaker verification model, this study aims to systematically evaluate state-of-the-art models by establishing a comprehensive benchmark for Turkish. Widely used speaker verification models from the literature are compared on the Turkish dataset under a common experimental protocol to identify the most success-

ful base model, thereby analyzing the zero-shot cross-lingual performance of models trained on different languages when applied to Turkish.

In the second stage, a comprehensive preprocessing analysis is conducted on the identified base model, evaluating methods such as normalization, pre-emphasis, windowing, noise reduction, and voice activity detection, both individually and in combination. The primary objective is to identify preprocessing strategies that reduce the False Accept (FA) rate while increasing similarity between reference and test audios, as well as to systematically examine how multi-stage filtering chains strengthen or weaken speaker-specific discriminative features.

2 Related Work

Speaker verification systems have made significant progress in text-independent tasks with the adoption of deep learning approaches. In particular, the transition from traditional vector-based methods to CNN and Transformer-based architectures has greatly increased the robustness of models under different acoustic conditions (Sang et al., 2023).

The difference between the language on which models are trained and the language on which they are tested brings about cross-lingual transfer problems. Recent studies in the literature systematically examining the cross-lingual transfer capabilities of different SV architectures show that model design plays a critical role in obtaining language-independent speaker representations (Buitrago and Hernando, 2026).

Besides model architectures, preprocessing techniques applied to speech signals also have a significant role on system performance. Various studies emphasize that steps such as voice activity detection (VAD), silence removal, noise reduction, and loudness normalization can positively affect the quality of the obtained embeddings and thus the verification performance (Labied et al., 2022).

The equal and fair performance of speaker verification systems across all demographic groups is one of the most critical issues that speech research has focused on in recent years. It reveals that system performance can show significant variations according to demographic characteristics such as age, gender, and dialect, and that this situation can lead to bias in models (Hutiri and Ding, 2022).

Large-scale benchmark datasets such as VoxCeleb play a central role in the advancement of

speaker verification research and the standardized comparison of different methods. These datasets offer consistent evaluation protocols to the community to measure the robustness of models under different recording conditions and populations (Nagrani et al., 2017). In the literature, a study focused on the effect of language mismatch in Turkish speaker verification using a dataset of 73 speakers (47 male, 26 female), showing that language mismatch between training and test data significantly degrades performance (Hanilçi, 2017).

3 Methodology

3.1 Dataset

In this study, a novel benchmark dataset was created for use in Turkish speaker verification research. The dataset was designed to offer a more comprehensive and distinct structure compared to existing speaker verification datasets in the literature in terms of demographic diversity and regional dialect coverage. Thus, it provides not only a benchmark for measuring the general verification performance of models but also enables the investigation of potential biases associated with performance variations across different demographic structures.

Category	Speakers	Recordings	%
Total	461	1,844	100
<i>Gender</i>			
Male	310	1,240	67
Female	151	604	33
<i>Age</i>			
Adult (< 50)	293	1,172	63.6
Elder (≥ 50)	168	672	36.4
<i>Regional Dialect</i>			
Istanbul	381	1,524	83
Black Sea	24	96	5
Central Anatolia	21	84	4
Eastern	19	76	4
Thrace	8	32	2
Aegean	8	32	2

Table 1: Dataset statistics.

TurkSV-Bench consists of a total of 1,844 distinct voice recordings belonging to 461 different speakers. Each speaker is represented by four fixed-length voice samples extracted from different and non-overlapping time intervals of the YouTube video from which the audio was sourced. Maintaining four recordings per speaker allows for the creation of multiple genuine verification pairs among different samples belonging to the same speaker and enables a more reliable measurement of intra-speaker variability.

In the construction of the dataset, gender, age, and regional dialect distributions were considered jointly to support subgroup-based evaluation and

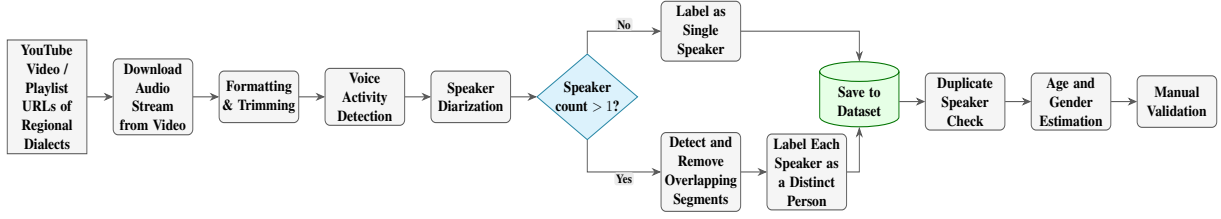


Figure 1: Pipeline for building the speaker verification benchmark dataset.

increase representational diversity. The dataset includes female and male speakers; adult, and elder age groups; and speakers representing Istanbul, Black Sea, Central Anatolia, Eastern, Thrace, and Aegean regional dialect categories. The resulting demographic and regional distribution is summarized in Table 1.

Field	Description
person_id	Speaker identity (P1, P2, ...)
data_id	Recording identity (P1_D1, ..., P1_D4)
url	Source video link
start / end	Start and end times in the video
person_number	Number of detected speakers
gender	Gender label
age_group	Age group (adult/elder)
region	Regional dialect label

Table 2: Metadatas stored for each audio recording.

Each recording is stored together with the metadata fields defined in Table 2.

The end-to-end pipeline followed in creating the benchmark dataset is shown in Figure 1. The pre-processing and quality control steps applied to the created dataset include the following:

Only the audio stream was extracted from the source videos; the recordings were converted to 16 kHz, mono, and WAV format. To increase the probability of excluding credits, music, advertisements, and promotional sections frequently found at the beginning of the videos, the first 60 seconds were not processed. The processing interval was limited to the 1000th second to ensure obtaining data of sufficient length from each person, considering the possibility of multiple people in a single video. Voice activity detection (VAD) was performed using Silero VAD (Silero Team, 2021), silent sections were removed, and only speech-containing regions were preserved. Speaker diarization was carried out with the pyannote.audio 3.1 pipeline (Plaquet and Bredin, 2023; Bredin, 2023) to determine the time intervals of the speakers, overlapping speech was cleaned, and each speaker was recorded with a separate identity. A 14-second segment was extracted from each recording; to eliminate overlaps caused by speaker transitions and clipped words concentrated at the edges, one second was further trimmed from each end, resulting in the final 12-

second recording.

To identify duplicate speakers semi-automatically, speaker embeddings extracted using an off-the-shelf ECAPA-TDNN model (Desplanques et al., 2020) trained on VoxCeleb were compared pairwise, and identities showing high similarity were manually examined and removed from the dataset. The age group and gender labels of the speakers were predicted using a wav2vec 2.0-based age and gender estimation model (Burkhardt et al., 2023). Following the automated processing stages, a manual quality-control procedure was applied to all recordings. Each speaker in TurkSV-Bench is represented by four 12-second recordings, and all four recordings associated with every speaker identity were listened to individually. This review was used to verify consistency among the recordings assigned to the same speaker and to identify residual errors that could remain after speaker diarization, overlap removal, and automatic metadata estimation. The automatically predicted age-group and gender labels were also subjected to human verification.

3.2 Models

The fundamental architectural features of current deep learning-based models, which are examined within the scope of the study and stand out in the literature for the speaker verification task, are summarized.

WeSpeaker ResNet34-LM is a ResNet34-based speaker verification model that extracts speaker-specific discriminative features from the audio signal. The model learns local patterns in the time and frequency axes by processing log-Mel filterbank features with two-dimensional convolutional layers. The model, which has approximately 6.63 million parameters, is trained on the VoxCeleb2-dev dataset, and large-margin fine-tuning is applied to the model to better separate different speakers in the embedding space (Wang et al., 2023a).

ECAPA-TDNN is a speaker verification model that improves the x-vector architecture with chan-

nel attention mechanisms and multi-scale feature aggregation. The model combines layer outputs at different depths using Res2Net-based blocks; it generates fixed-dimensional embeddings from variable-length audios through attentive statistics pooling. The model, with approximately 20.8 million parameters, is trained on the VoxCeleb1 and VoxCeleb2 datasets (Desplanques et al., 2020).

CAM++ is a densely connected TDNN (D-TDNN) based speaker verification model developed to extract speaker-specific information. The model processes time-frequency details with a two-dimensional convolutional front-end and aims to highlight speaker-related features and reduce the effect of noise through a context-aware masking mechanism. The model, which has a multi-scale pooling structure, contains approximately 7.18 million parameters (Wang et al., 2023b).

The ERes2Net family consists of speaker verification models that enhance the multi-scale structure of Res2Net with attention-based feature fusion mechanisms. The base ERes2Net aims to produce more discriminative speaker embeddings by combining local details within residual blocks and information from different network stages. ERes2NetV2 is a model with 17.8 million parameters that increases computational efficiency with fusion components that simplify the structure (Chen et al., 2024). ERes2Net-large is the higher-capacity version of the base structure with 22.46 million parameters (Chen et al., 2023).

TitaNet-Large is a ContextNet-based model developed for speaker verification. The model learns local and global speaker-specific features by combining one-dimensional depthwise separable convolutions with Squeeze-and-Excitation layers; it transforms variable-length recordings into 192-dimensional t-vector embeddings using channel-attentive statistical pooling. It is a model with approximately 23 million parameters (Koluguri et al., 2022).

ESPnet ECAPA-WavLM combines the self-supervised speech representations of WavLM-Large with an ECAPA-TDNN speaker encoder: the WavLM front-end extracts contextual and speaker-specific information from raw audio, which ECAPA-TDNN transforms into discriminative speaker embeddings. The model is trained on the VoxCeleb1 and VoxCeleb2 development sets (Jung et al., 2024).

WavLM-MHFA combines the self-supervised speech representations of WavLM-Base+ with a

Model	TO-HTER	ROCCH-		FAR@		TO-HTER		ROCCH-	
		EER	minDCF	FRR=0	FA FR	E[FA]	E[FR]		
ResNet34-LM	0.0055	0.0107	0.0091	0.0110	186 0	181.3	0.30		
ERes2NetV2	0.0598	0.0635	0.0122	0.1812	801 2	1077.5	1.76		
ECAPA-TDNN	0.0631	0.0652	0.0146	0.1676	915 2	1105.5	1.80		
ERes2Net	0.0732	0.0612	0.0117	1.1476	1256 2	1037.9	1.69		
TitaNet-Large	0.0812	0.0591	0.0100	0.2045	1529 2	1002.2	1.63		
ERes2Net-Large	0.1802	0.1659	0.0226	0.8196	3048 5	2814.6	4.59		
WavLM-MHFA	0.1803	0.1714	0.0153	1.3176	3052 5	2908.5	4.74		
ECAPA-WavLM	0.2227	0.2137	0.0193	1.5746	3876 6	3625.6	5.91		
CAM++	0.2807	0.2470	0.0213	3.4263	4618 8	4190.0	6.83		

Table 3: Benchmark results for speaker verification models on TurkSV-Bench.

Multi-Head Factorized Attentive Pooling (MHFA) layer, which fuses information from different Transformer layers via multiple attention heads to enhance speaker discrimination. With its 64-head structure, the model contains 96.2 million parameters (Peng et al., 2023).

3.3 Preprocessing Methods

Prior to the embedding extraction stage, various preprocessing steps described below are performed in order to reduce systemic or environmental distorting effects in the speech signals and to make the recordings optimally suited for the network architecture.

Peak Normalization: All samples are scaled to a specific range by taking the absolute highest amplitude value of the signal as a reference. This fundamental operation aims to balance the general volume differences among different audio recordings.

RMS Normalization: The root mean square energy of the signal is calculated and adjusted to a target level. Since it is based on the average energy rather than peak values, it provides a more robust and consistent normalization against sudden amplitude spikes in the speech signal.

Amplitude Scaling: Preserving the original morphological features of the signal, the amplitude values are multiplied linearly by a constant coefficient. It is a fundamental operation that only adjusts the magnitude of the signal and does not cause structural distortion.

Loudness Normalization (LUFS): In this method, the signal is brought to a perceptual target loudness level (-23 LUFS in this study) by taking the frequency sensitivity of the human auditory system into account. Unlike energy-based approaches, it aims to increase perceptual consistency.

DC Offset Removal: By bringing the arithmetic mean of the signal to zero, DC offsets that may originate from recording devices or the audio chain are cleaned. Thus, low-frequency artifact compo-

nents are prevented from negatively affecting the feature extraction stage.

Pre-emphasis: A first-order high-pass filter is applied to balance the natural spectral tilt of the speech signal. This operation makes speaker-specific formant information in higher frequency bands more prominent. In the study, the values 0.95, 0.97, and 0.99 are examined comparatively to analyze the effect of the α coefficient on performance.

Cepstral Mean Normalization (CMN): Used to suppress static spectral distortions from the channel and recording environment. Since models use end-to-end feature extraction, a waveform-level equivalent is designed by resynthesizing the signal after subtracting the band-based logarithmic amplitude mean in the STFT domain.

Windowing: Hamming, Hanning, and Blackman window functions are tested to minimize spectral leakage and smooth discontinuities at frame boundaries. The signal is windowed by dividing it into frames of 25 ms length with a 10 ms shift, and then the waveform is reconstructed using the overlap-add method.

Spectral Subtraction: The noise spectrum estimated from the silent (low-energy) sections of the recording is subtracted from the overall amplitude spectrum of the signal. In this way, stationary background noise is suppressed, ensuring that speaker features are obtained more clearly.

Wiener Filtering: An adaptive filtering is performed with the Minimum Mean Square Error (MMSE) criterion, based on the statistical properties of the signal and noise. This method aims to directly improve the speaker verification performance by increasing the Signal-to-Noise Ratio (SNR).

4 Experiments

The experimental studies were conducted on the TurkSV-Bench dataset detailed in Section 3.1, comprising 1,844 speech recordings from 461 speakers. A total of 1,699,246 evaluation trials were constructed, consisting of 2,766 genuine and 1,696,480 impostor trials.

System performance was evaluated using Threshold-Optimized Half Total Error Rate (TO-HTER), ROC Convex Hull Equal Error Rate (ROCCH-EER), the minimum Detection Cost Function (minDCF), and FAR@FRR=0. False Accept (FA) and False Reject (FR) counts were ad-

ditionally reported at the corresponding operating points.

Cosine similarity between the speaker embedding vectors extracted from each pair of speech samples was used as the verification score. For a decision threshold θ , a pair was accepted as belonging to the same speaker when its cosine similarity score was greater than or equal to θ , and rejected otherwise. Accordingly, a false accept (FA) occurs when an impostor pair is incorrectly accepted, whereas a false reject (FR) occurs when a genuine pair is incorrectly rejected. FAR and FRR were calculated as the percentages of FA and FR relative to the total number of impostor and genuine pairs, respectively, at each decision threshold. We additionally report FAR@FRR=0, defined as the false acceptance rate at the strictest empirically attainable decision threshold for which no genuine trial is falsely rejected.

At a decision threshold θ , HTER is computed as the arithmetic mean of FAR(θ) and FRR(θ) (Bengio and Mariéthoz, 2004). The empirical TO-HTER operating point is selected by minimizing $|FAR(\theta) - FRR(\theta)|$ over the available evaluation thresholds.

Equal-error performance is additionally characterized using ROCCH-EER, obtained from the intersection of the ROC convex hull with the FAR = FRR diagonal (Brümmer and de Villiers, 2013). This formulation provides an equal-error estimate when the discrete empirical operating points do not contain an exact equality between FAR and FRR, which is particularly relevant in low-error operating regions.

The minDCF is additionally reported as the minimum application-weighted detection cost over the evaluated decision thresholds.

Observed errors at TO-HTER are denoted as FA@TO-HTER and FR@TO-HTER. Expected errors at ROCCH-EER are reported as E[FA]@ROCCH-EER and E[FR]@ROCCH-EER, which can be non-integers since this operating point may lie between empirically attainable points.

The evaluation consists of two stages. In the first stage, the speaker verification models introduced in Section 3.2 were compared to identify the architecture yielding the most discriminative representation, establishing it as the baseline. For this stage, overall results computed across the entire dataset are presented in Table 3, while performance broken down by gender, age, and regional dialect

is reported in Table 4.

In the second stage, using this baseline model, we analyzed the impact of the signal preprocessing techniques described in Section 3.3 on verification performance. Similarly, overall results computed across the entire dataset are presented in Table 5, while the corresponding breakdown by gender, age, and regional dialect is reported in Table 6.

As part of the preprocessing analysis, all possible pairwise and triple-wise chain filter combinations of the individual filters examined in this study were tested. Due to space constraints and readability, the results of all combinations could not be presented in detail; only the most successful filter chains are reported in this paper.

5 Results

The evaluated pretrained speaker verification systems demonstrated varying levels of zero-shot generalization to Turkish speech. Among the models reported in Table 3, ResNet34-LM achieved the strongest overall performance, with a TO-HTER of 0.0055%, a ROCCH-EER of 0.0107%, and a minDCF of 0.0091. ResNet34-LM was therefore used as the baseline architecture for the subsequent preprocessing analyses.

The demographic subgroup results in Table 4 show relatively small differences between the male and female subsets for ResNet34-LM. The model yielded TO-HTER values of 0.0100% for male speakers and 0.0080% for female speakers, with corresponding ROCCH-EER values of 0.0187% and 0.0152%, respectively. For the Adult and Elder age groups reported in the subgroup comparison, TO-HTER values were 0.0076% and 0.0116%, while ROCCH-EER values were 0.0148% and 0.0217%, respectively. A larger separation was observed across regional-dialect groups. The Istanbul subset yielded a TO-HTER of 0.0057% and a ROCCH-EER of 0.0112%, whereas the Other Dialects subset yielded 0.0356% and 0.0656%, respectively.

Table 5 summarizes the overall preprocessing results obtained with ResNet34-LM. Among the individual preprocessing methods, pre-emphasis with $\alpha = 0.97$ yielded a TO-HTER of 0.0051% and a ROCCH-EER of 0.0096%, while CMN yielded 0.0052% and 0.0098%, respectively. The selected pairwise and triple-filter configurations remained within a similarly narrow low-error range. In contrast, spectral subtraction and Wiener filtering pro-

duced substantially higher error rates across the reported evaluation metrics.

The subgroup preprocessing results in Table 6 further show that the effect of preprocessing depends on both subgroup and evaluation metric. For male speakers, the baseline TO-HTER of 0.0100% changed to 0.0078% with pre-emphasis at $\alpha = 0.99$, while for female speakers the baseline value of 0.0080% changed to 0.0077% with pre-emphasis at $\alpha = 0.97$. For Adult speakers, the baseline TO-HTER of 0.0076% and ROCCH-EER of 0.0148% decreased to 0.0075% and 0.0142%, respectively, with either CMN or pre-emphasis at $\alpha = 0.97$. Pre-emphasis at $\alpha = 0.99$ reduced Adult minDCF from 0.0129 to 0.0117, but increased TO-HTER to 0.0080%. For Elder speakers, the baseline TO-HTER of 0.0116% changed to 0.0082% with either CMN or pre-emphasis at $\alpha = 0.99$. For the Other Dialects subset, the baseline TO-HTER of 0.0356% changed to 0.0297% with CMN and to 0.0287% for the reported Blackman+CMN-based pairwise and triple-filter configurations. These changes reflect the effects observed at the corresponding operating points, which varied across the evaluated preprocessing configurations. Spectral subtraction and Wiener filtering, in contrast, consistently produced substantially higher error rates across the demographic and regional-dialect subsets.

6 Discussion and Conclusion

In this study, we present an end-to-end data generation pipeline that automates the process from data collection to quality control. Using this infrastructure, we benchmark a diverse set of pretrained speaker-verification models, evaluating their zero-shot generalization to Turkish speech as well as the effects of several speech preprocessing methods.

The model comparison shows that performance on Turkish speech varies substantially across architectures. ResNet34-LM achieved the strongest overall results despite having considerably fewer parameters than several competing systems. This suggests that model capacity alone does not determine cross-lingual speaker-verification performance and that architectural design and training strategy may play an important role in generalization to Turkish speech.

From a demographic perspective, gender- and age-related differences were relatively limited and varied across models. In contrast, more noticeable differences were observed between the Istanbul

and Other Dialects subsets, with the Istanbul subset generally yielding lower error rates across the evaluated systems.

The preprocessing experiments show that most methods produced only small changes in overall verification performance. The observed improvements were limited to specific preprocessing methods and evaluation metrics, while the results varied across demographic subgroups. Pre-emphasis, CMN, and several filter combinations produced

small improvements in some evaluation metrics. These findings indicate that the effect of preprocessing was dependent on the evaluation setting, with some configurations providing modest improvements for certain demographic subgroups. In contrast, spectral subtraction and Wiener filtering resulted in consistently higher error rates, indicating that these more intensive noise-suppression methods were less suitable for the speaker-recognition conditions considered in this study.

(a) Gender												
Model	Male					Female						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0100	0.0187	0.0166	0.0200	153 0	143.59 0.35	0.0080	0.0152	0.0066	0.0160	29 0	27.55 0.14
ERes2NetV2	0.0628	0.0402	0.0153	0.0718	550 1	308.30 0.75	0.2707	0.1727	0.0110	0.5359	581 2	312.97 1.56
ECAPA-TDNN	0.0550	0.0463	0.0200	0.0826	431 1	354.82 0.86	0.2889	0.1641	0.0154	0.5033	647 2	297.31 1.49
ERes2Net	0.0579	0.0591	0.0170	0.1254	475 1	452.81 1.10	0.1209	0.1218	0.0110	0.2949	238 1	220.77 1.10
TitaNet-Large	0.0838	0.0875	0.0158	0.3003	460 2	670.21 1.63	0.0952	0.0511	0.0110	0.0800	145 1	92.65 0.46
ERes2Net-Large	0.2183	0.2025	0.0273	0.5280	1698 4	1551.48 3.77	0.3237	0.2581	0.0270	1.4134	573 3	467.71 2.34
WavLM-MHFA	0.1615	0.1327	0.0200	1.0041	1239 3	1016.78 2.47	0.4821	0.3966	0.0225	4.0121	947 4	718.63 3.59
ECAPA-WavLM	0.2190	0.2014	0.0225	2.2620	1709 4	1543.65 3.75	0.4818	0.3710	0.0368	4.5817	946 4	672.31 3.36
CAM++	0.3747	0.3132	0.0253	5.7355	2859 7	2400.19 5.83	0.4418	0.3469	0.0199	1.0822	801 4	628.64 3.14

(b) Age												
Model	Adult (< 50)					Elder (≥ 50)						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0076	0.0148	0.0129	0.0152	104 0	101.02 0.26	0.0116	0.0217	0.0148	0.0232	52 0	48.72 0.22
ERes2NetV2	0.1146	0.0785	0.0106	0.1921	790 2	537.32 1.38	0.1026	0.0689	0.0277	0.1060	238 1	154.65 0.69
ECAPA-TDNN	0.1268	0.0895	0.0191	0.1893	957 2	612.47 1.57	0.0172	0.0320	0.0180	0.0343	77 0	71.84 0.32
ERes2Net	0.0666	0.0699	0.0113	1.0359	523 1	478.39 1.23	0.0964	0.0668	0.0280	0.0936	210 1	150.03 0.67
TitaNet-Large	0.1239	0.0854	0.0150	0.2376	918 2	584.74 1.50	0.0080	0.0158	0.0146	0.0160	36 0	35.52 0.16
ERes2Net-Large	0.1188	0.1206	0.0214	0.2444	847 2	825.64 2.12	0.2767	0.2858	0.0362	1.6717	574 3	641.46 2.88
WavLM-MHFA	0.2908	0.2579	0.0227	1.6542	2034 5	1765.10 4.53	0.0207	0.0354	0.0200	0.0414	93 0	79.44 0.36
ECAPA-WavLM	0.3395	0.3241	0.0285	1.9150	2311 6	2218.11 5.70	0.0187	0.0354	0.0235	0.0374	84 0	79.37 0.36
CAM++	0.1856	0.1353	0.0188	0.5103	1373 3	926.01 2.38	0.5939	0.5119	0.0333	6.5663	1330 6	1149.06 5.16

(c) Regional Dialect												
Model	Istanbul					Other Dialects						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0057	0.0112	0.0098	0.0114	132 0	129.28 0.26	0.0356	0.0656	0.0397	0.0712	36 0	33.17 0.31
ERes2NetV2	0.0486	0.0312	0.0068	0.0534	618 1	361.71 0.71	0.5149	0.2908	0.0703	0.6290	310 2	147.02 1.40
ECAPA-TDNN	0.0411	0.0400	0.0110	0.1874	445 1	463.41 0.91	0.3923	0.2926	0.0592	0.6052	186 2	147.96 1.40
ERes2Net	0.0055	0.0107	0.0065	0.0111	128 0	123.84 0.24	0.4170	0.3423	0.0731	4.4086	211 2	173.06 1.64
TitaNet-Large	0.0934	0.0690	0.0109	0.2261	1150 2	799.39 1.58	0.2031	0.1175	0.0361	0.1978	100 1	59.40 0.56
ERes2Net-Large	0.0912	0.0869	0.0170	0.2314	1100 2	1007.04 1.99	0.6566	0.6148	0.0779	1.7880	348 3	310.84 2.95
WavLM-MHFA	0.1237	0.1216	0.0159	0.4029	1345 3	1408.25 2.78	0.7812	0.4998	0.0460	1.7820	474 3	252.69 2.40
ECAPA-WavLM	0.1317	0.0996	0.0194	0.2471	1531 3	1153.30 2.28	1.0005	0.7736	0.0522	1.8097	485 5	391.11 3.71
CAM++	0.1996	0.1561	0.0132	3.7873	2596 4	1807.83 3.57	0.8785	0.8067	0.0758	1.6594	467 4	407.89 3.87

Table 4: Benchmark results of models across (a) gender, (b) age, and (c) regional dialect subgroups on TurkSV-Bench.

Preprocessing Method	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
Baseline	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Pre-emphasis ($\alpha = 0.99$)	0.0053	0.0098	0.0079	0.0106	180 0	166.43 0.27
CMN	0.0052	0.0098	0.0084	0.0104	176 0	167.01 0.27
Pre-emphasis ($\alpha = 0.97$)	0.0051	0.0096	0.0084	0.0102	173 0	163.41 0.27
Amplitude scaling	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
RMS normalization	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
DC offset removal	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Peak normalization	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Loudness normalization (-23 LUFS)	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Hamming windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Hann windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.37 0.30
Blackman windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.37 0.30
Spectral subtraction	1.0113	0.9782	0.1617	11.0728	17140 28	16594.97 27.06
Wiener filtering	2.0258	2.0219	0.1856	31.1352	34388 56	34301.13 55.93
Pairwise filter combinations						
Blackman + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27
Hann + Pre-emphasis ($\alpha = 0.97$)	0.0051	0.0096	0.0084	0.0101	172 0	162.71 0.27
Triple filter combinations						
Blackman + Hann + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27
RMS normalization + Hann + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27

Table 5: Benchmark results of preprocessing methods applied with the ResNet34-LM model.

(a) Age												
Preprocessing Method	Adult (< 50)						Elder (≥ 50)					
	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
Baseline	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Peak normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
RMS normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Amplitude scaling	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
CMN	0.0075	0.0121	0.0142	0.0150	103 0	97.25 0.25	0.0082	0.0136	0.0160	0.0165	37 0	35.87 0.16
Loudness normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
DC offset removal	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Pre-emphasis ($\alpha = 0.97$)	0.0075	0.0120	0.0142	0.0150	103 0	97.25 0.25	0.0085	0.0135	0.0163	0.0169	38 0	36.68 0.16
Pre-emphasis ($\alpha = 0.99$)	0.0080	0.0117	0.0147	0.0159	109 0	100.72 0.26	0.0082	0.0135	0.0160	0.0165	37 0	35.87 0.16
Hamming windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Hann windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Blackman windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Spectral subtraction	0.8581	0.1792	0.8424	4.5529	5907 15	5765.66 14.81	1.1903	0.1503	1.1303	10.5579	2671 12	2536.86 11.39
Wiener filtering	2.2054	0.2042	2.1762	29.4458	15005 39	14894.92 38.26	2.0893	0.1902	1.9439	14.0215	4703 21	4363.13 19.59
Pairwise filter combinations												
Blackman + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16
Hann + Pre-emphasis ($\alpha = 0.97$)	0.0075	0.0120	0.0142	0.0150	103 0	97.25 0.25	0.0085	0.0135	0.0163	0.0169	38 0	36.68 0.16
Triple filter combinations												
Blackman + Hann + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16
RMS + Hann + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16

(b) Gender												
Preprocessing Method	Male						Female					
	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
Baseline	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Peak normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
RMS normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Amplitude scaling	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
CMN	0.0083	0.0159	0.0163	0.0166	127 0	124.58 0.30	0.0086	0.0071	0.0156	0.0171	31 0	28.18 0.14
Loudness normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
DC offset removal	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Pre-emphasis ($\alpha = 0.97$)	0.0082	0.0158	0.0161	0.0163	125 0	123.35 0.30	0.0077	0.0055	0.0142	0.0155	28 0	25.69 0.13
Pre-emphasis ($\alpha = 0.99$)	0.0078	0.0155	0.0155	0.0157	120 0	118.94 0.29	0.0088	0.0055	0.0159	0.0177	32 0	28.83 0.14
Hamming windowing	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Hann windowing	0.0100	0.0166	0.0188	0.0200	153 0	143.92 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Blackman windowing	0.0100	0.0166	0.0188	0.0200	153 0	143.92 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Spectral subtraction	1.2311	0.1897	1.1736	13.6497	9392 23	8993.42 21.83	1.5494	0.2173	1.5007	27.2853	2815 14	2719.30 13.60
Wiener filtering	1.7192	0.1648	1.7046	47.2108	13165 32	13063.00 31.71	3.5055	0.2611	3.4075	47.1181	6304 32	6174.39 30.87
Pairwise filter combinations												
Blackman + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13
Hann + Pre-emphasis ($\alpha = 0.97$)	0.0082	0.0157	0.0161	0.0163	125 0	123.35 0.30	0.0077	0.0055	0.0142	0.0155	28 0	25.69 0.13
Triple filter combinations												
Blackman + Hann + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13
RMS + Hann + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13

(c) Regional Dialect												
Preprocessing Method	Istanbul						Other Dialects					
	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	minDCF	ROCCH-EER	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
Baseline	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Peak normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
RMS normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Amplitude scaling	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
CMN	0.0057	0.0093	0.0108	0.0113	131 0	125.31 0.25	0.0297	0.0441	0.0566	0.0593	30 0	28.64 0.27
Loudness normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
DC offset removal	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Pre-emphasis ($\alpha = 0.97$)	0.0056	0.0089	0.0105	0.0111	129 0	122.02 0.24	0.0356	0.0402	0.0668	0.0712	36 0	33.76 0.32
Pre-emphasis ($\alpha = 0.99$)	0.0058	0.0086	0.0108	0.0117	135 0	125.12 0.25	0.0346	0.0402	0.0655	0.0692	35 0	33.11 0.31
Hamming windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Hann windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0346	0.0397	0.0643	0.0692	35 0	32.53 0.31
Blackman windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0346	0.0397	0.0643	0.0692	35 0	32.53 0.31
Spectral subtraction	0.9129	0.1611	0.8918	11.1952	10508 21	10329.61 20.39	0.6141	0.1217	0.5698	4.4165	305 3	288.09 2.74
Wiener filtering	2.2315	0.1940	2.1633	30.9427	25853 51	25056.17 49.45	2.0405	0.2101	1.9488	20.6112	1010 10	985.32 9.35
Pairwise filter combinations												
Blackman + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27
Hann + Pre-emphasis ($\alpha = 0.97$)	0.0055	0.0089	0.0105	0.0111	128 0	121.30 0.24	0.0356	0.0402	0.0668	0.0712	36 0	33.76 0.32
Triple filter combinations												
Blackman + Hann + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27
RMS + Hann + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27

Table 6: Preprocessing results across (a) age, (b) gender, and (c) regional dialect subgroups.

Limitations

Due to the nature of the data collection platform, TurkSV-Bench exhibits partial imbalances in the distribution of age, gender, and regional dialects. Future work can extend the benchmark with larger and more balanced subgroup distributions, and additional acoustic conditions to provide more robust estimates of real-world speaker-verification perfor-

mance. Additionally, as both reference and test recordings are extracted from the same source, the resulting performance metrics may be somewhat optimistic compared to true cross-session conditions.

Ethics Statement

The audio recordings collected within the scope of TurkSV-Bench have been anonymized and uti-

lized solely for research purposes. Due to personal data usage policies, raw audio files cannot be shared. However, the data collection methodology and metadata are provided in detail to researchers.

References

- Samy Bengio and Johnny Mariéthoz. 2004. [A statistical significance test for person authentication](#). In *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*.
- Hervé Bredin. 2023. [pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe](#). In *Proc. INTERSPEECH 2023*, pages 1983–1987.
- Niko Brümmer and Edward de Villiers. 2013. [The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF](#). *arXiv preprint arXiv:1304.2865*.
- Pol Buitrago and Javier Hernando. 2026. [Towards language-agnostic speaker verification: A cross-lingual transfer study of architectures](#). In *Proc. Odyssey 2026*.
- Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn W. Schuller. 2023. [Speech-based age and gender prediction with transformers](#). In *Speech Communication – 15th ITG Conference*, pages 46–50, Aachen, Germany. VDE.
- Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023. [An enhanced Res2Net with local and global feature fusion for speaker verification](#). In *Proc. INTERSPEECH 2023*, pages 2228–2232.
- Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. 2024. [ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency](#). In *Proc. INTERSPEECH 2024*, pages 3245–3249.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification](#). In *Proc. Interspeech 2020*, pages 3830–3834.
- Yue Fan, J. W. Kang, L. T. Li, K. C. Li, H. L. Chen, S. T. Cheng, P. Y. Zhang, Z. Y. Zhou, Y. Q. Cai, and Dong Wang. 2020. Cn-celeb: A challenging chinese speaker recognition dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608.
- Cemal Hanilçi. 2017. [Effect of language mismatch on turkish speaker verification](#). *Uludağ University Journal of The Faculty of Engineering*, 22(1):189–196.
- Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. [Bias in automated speaker recognition](#). In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 230–247.
- Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Alex Gichamba, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe. 2024. [ESPnet-SPK: Full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models](#). In *Proc. INTERSPEECH 2024*.
- Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. [TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context](#). In *Proc. IEEE ICASSP 2022*.
- Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. 2022. [An overview of automatic speech recognition preprocessing techniques](#). In *Proc. 2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pages 804–809.
- Arsha Nagrani, Joon Son Chung, and Andrew Senior. 2017. [VoxCeleb: a large-scale speaker identification dataset](#). In *Proc. INTERSPEECH 2017*, pages 2616–2620.
- Junyi Peng, Oldřich Plchot, Themis Stafylakis, Ladislav Mošner, Lukáš Burget, and Jan Černocký. 2023. [An attention-based backend allowing efficient fine-tuning of Transformer models for speaker verification](#). In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.
- Alexis Plaquet and Hervé Bredin. 2023. [Powerset multi-class cross entropy loss for neural speaker diarization](#). In *Proc. INTERSPEECH 2023*, pages 3222–3226.
- Mufan Sang, Yong Zhao, Gang Liu, John H. L. Hansen, and Jian Wu. 2023. [Improving Transformer-based networks with locality for automatic speaker verification](#). In *Proc. ICASSP 2023*, pages 1–5.
- Silero Team. 2021. Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. <https://github.com/snakers4/silero-vad>.
- Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023a. [WeSpeaker: A research and production oriented speaker embedding learning toolkit](#). In *Proc. IEEE ICASSP 2023*.
- Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. [CAM++: A fast and efficient network for speaker verification using context-aware masking](#). In *Proc. INTERSPEECH 2023*, pages 5301–5305.