IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/44AI/2024/4441424/
Vol. 2024 (22), Article ID 4441424

Expanding The Functionality of Speech Corpora for Broader Applications*

Michal HALON and Andrzej PACUT

NASK National Research Institute, Warsaw, Poland

Correspondence should be addressed to: Michal HALON, michal.halon@nask.pl

* Presented at the 44th IBIMA International Conference, 27-28 November 2024 Granada, Spain

Abstract

Numerous speech corpora have been developed for a variety of purposes. While some are designed for specific topics, they often can be adapted for broader applications, especially when suitable datasets for specific domains or languages are scarce. This paper introduces a method for adapting existing speech corpora for use in biometric recognition and personalized speech synthesis. This involves verifying the accuracy of key metadata in the processed dataset, such as gender labeling and speaker attribution. To accomplish this, we propose a method that analyzes biometric verification distributions of voice samples. Potential inaccuracies are flagged for subsequent human expert listening analysis. The method was tested on the Clarin-PL polish speech corpus using Phonexia software, resulting in improved biometric recognition metrics, including Equal Error Rate (EER) and False Acceptance/False Rejection Rates (FAR/FRR). Our findings demonstrate that this approach can significantly enhance the reliability and applicability of speech corpora in extended applications, especially for those where suitable datasets are scarce. By reducing the need for extensive manual verification, the proposed method facilitates broader utilization of existing speech corpora for advanced biometric and speech synthesis tasks.

Keywords: speech corpus, voice biometrics, speech synthesis

Introduction

Biometric recognition is becoming increasingly integral to our daily lives, with voice biometrics emerging as one of key identity verification method (González-Rodríguez et al., 2008). This modality has spurred the development of various practical applications and algorithms (Campbell, 1997; Chung et al., 2018; Chakroun and Frikha, 2023).

Human speech generation has made remarkable progress, utilizing advanced machine learning algorithms to create natural and fluent synthetic voices. This technology is applied in virtual assistants, accessibility tools, and content creation, enhancing user interactions and enabling personalized communication (Tan et al., 2021; Budzianowski et al., 2024).

Developing and testing such a systems necessitates access to a speech corpus with correctly prepared and labeled speaker recordings. However, access to such corpora, especially in non-English languages, is limited. Furthermore,

Cite this Article as: Michal HALON and Andrzej PACUT, Vol. 2024 (22) "Expanding The Functionality of Speech Corpora for Broader Applications " Communications of International Proceedings, Vol. 2024 (22), Article ID 4441424, https://doi.org/10.5171/2024.4441424

not all databases are properly labeled, and some are not designed for intended tasks, leading to potential issues. Existing inaccuracies in these corpora can hinder effective utilization. While studies have addressed annotation errors in contexts like text-to-speech or speech recognition (Matoušek and Tihelka, 2013; Chiari, 2007), research on error detection in voice biometrics corpora is generally lacking. In this paper we outline steps to identify and correct annotation inaccuracies in speech corpora metadata, specifically for the development and testing of biometric verification and identification software, as well as for another rapidly growing in recent years - human speech generation field.

Speech corpora for biometric recognition and human speech generation

A speech corpus for biometric or human speech generation system development and testing must meet specific criteria. Firstly, it should include a diverse array of speakers and provide sufficiently lengthy recordings for each. Secondly, accurate labeling is crucial: recordings must be correctly attributed to the respective speakers, and to comply with common no-cross-gender comparison standards in voice biometrics testing (NIST, 2018), gender labeling of each speaker is necessary. Thirdly, the corpus should detail the recording conditions (e.g., studio-quality audio, phone recordings with background noise) and specify the recording equipment used.

Inaccuracies in speech corpora can significantly impact the performance and accuracy of biometric or human speech generation systems. During our research, we identified four primary types of errors affecting such systems. These include (Fig 1):

- (e1) Gender mislabeling,
- (e2) Incorrectly assigning recordings of one speaker to other speakers,
- (e3) Allocating recordings from multiple speakers to a single speaker's directory,
- (e4) Mistakenly labeling different origins for recordings from the same speaker.

It's important to note that these errors, often human-made, can occur during the dataset creation, subsequent labeling or corpora adaptation processes. Examples of situations and human behaviors leading to these errors include:

- Misidentifying the number of speakers in a recording session (for example, indicating one speaker in a session instead of two),
- Unintentionally recording or assigning a non-primary speaker in a session (leading to the unmarked presence of speaker in session),
- Determining or entering a speaker's gender errors,
- Assigning incorrect speaker IDs or saving session recordings to the wrong directory (for example to existing session folder instead of creating a new one).

In this study, we hypothesize that human speech corpora would be error-free without the presence of the human errors described above.

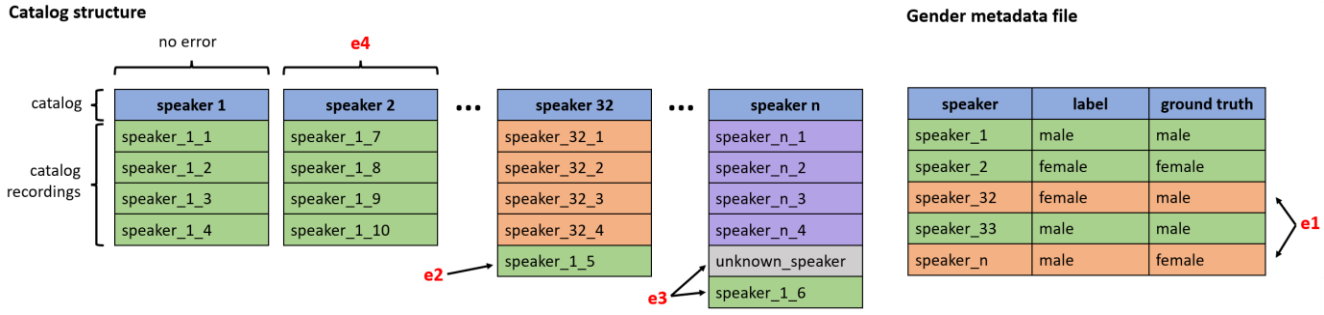

Fig 1. Visualization of detectable inaccuracies in speech corpora. Pictured on the left is the structure of folders containing recordings, highlighting incorrect placement of recordings and misnaming of folders as e2, e3, and e4 errors. On the right is an example of mislabeling gender assigned to speakers (e1).

Proposed approach

The process of verifying the accuracy of metadata essential for biometric recognition (as well as for other applications, such as generating human speech), can be done through listening. However, this method is time-consuming and error-prone for large datasets due to factors like listener fatigue. This chapter introduces a semi-automated method to identify and correct labeling inaccuracies in speech corpora. Semi-automated method means that the software automatically identifies problematic recordings for listening verification, and the number of such recordings should be only a fraction of the entire dataset content (depending on the degree of errors present in corpora), significantly speeding up the entire process. We applied this methodology to the Clarin-PL Studio Corpus (Korzinek et al., 2017) using Phonexia software (Phonexia, n.d.). The inaccuracies identification process has five steps:

- Analyzing genuine scores to find inaccuracies in the same speaker's recordings,
- Detecting gender labeling errors through no-cross-gender impostor score analysis,
- Searching for the same speakers using impostor score analysis,
- Deleting identified inaccuracies, including specific recordings and folders,
- Extensively analyzing impostor scores to verify the modifications.

All inaccuracies detected by the algorithm, including gender and recording issues, were confirmed through human expert listening analysis. For every listening analysis an appropriate sorted and labeled list of recordings (or pairs of recordings) was provided, which led to fast and convenient metadata verification.

Clarin-PL Studio Corpus (EMU)

The Clarin-PL speech corpus, recorded in a studio environment, utilizes two types of microphones: a high-quality studio microphone and a standard consumer audio headset. It comprises approximately 56 hours of Polish speech from 317 speakers across 554 sessions, each session containing 20 read sequences and 10 phonetically rich words (Korzinek et al., 2017). Designed for applications like speech-to-text, voice activity detection, speaker diarization, and keyword spotting, this corpus is available under an open license for both commercial and non-commercial use. The accessible version includes recordings from 295 speakers grouped in 553 sessions.

Originally, the database was organized by recording sessions, with each folder containing one session's recordings and additional information, including the speaker's ID and gender. For voice biometrics and human speech generation purposes, the storage format was modified: recordings from each speaker were consolidated into individual folders named after the speaker's ID. This reorganization, facilitated by an existing text file, resulted in 13,802 audio files being regrouped into 295 folders (instead of 553 session folders).

Phonexia software

Phonexia software provides voice solutions like language and gender identification, speech transcription, and voice activity detection. The Phonexia Speaker Identification (SID) module was utilized to implement and test speech corpora errors detection algorithm presented in this paper. This module identifies speakers regardless of language, accent, text, or channel. The outcome of voice print comparisons is given as a log-likelihood ratio (Phonexia, n.d.; Jessen et al., 2019). According to Phonexia, the recommended speech length for enrollment is at least 20 seconds and for identification/verification minimum 3 seconds (Phonexia, n.d.). However, biometric recognition in practice can be performed with samples as short as 2 seconds.

Metadata verification in recordings of the same speaker

To identify folders with recordings from multiple speakers, we performed a genuine scores analysis. We selected a representative recording for each speaker (it can be a random selection or first recording for each directory) and compared it against other recordings within the same speaker's folder using Phonexia software. The collective verification results were visualized in a data distribution plot (Fig 2, left side). The distribution was asymmetric and disrupted in negative direction, indicating incorrectly placed recordings. Based on this distribution, we set a biometric verification threshold. Recordings below this threshold underwent further processing. A second, lower threshold was also established. Recordings below this were automatically deemed inappropriate and deleted. Listening analysis confirmed this for all recordings below the second threshold. Recordings scoring between the two thresholds underwent comparative listening analysis against the folder's representative recording. Incompatible recordings were removed. If over 90% of files in a folder were deleted, the entire folder was removed from the corpus. Fig 2 (right side) shows the genuine scores distribution after removing mislabeled recordings.

The above two thresholds were determined based on visual assessment of the distribution of genuine scores. To make sure they were selected correctly, some of the samples were analyzed on the left (smaller threshold) and on the right (larger threshold) - all analyzed samples to the left of the first threshold had incorrect annotations, and all samples to the right of the second threshold were correct, which may indicate that there are actually mixed samples between the thresholds that require further listening analysis.

We chose the type of analysis described above because misplaced recordings compared to correctly placed files represent impostor comparisons. These scores should significantly differ from genuine scores, facilitating separation. Identified mislabeled recordings were further analyzed to find their original speaker and relocate them correctly.

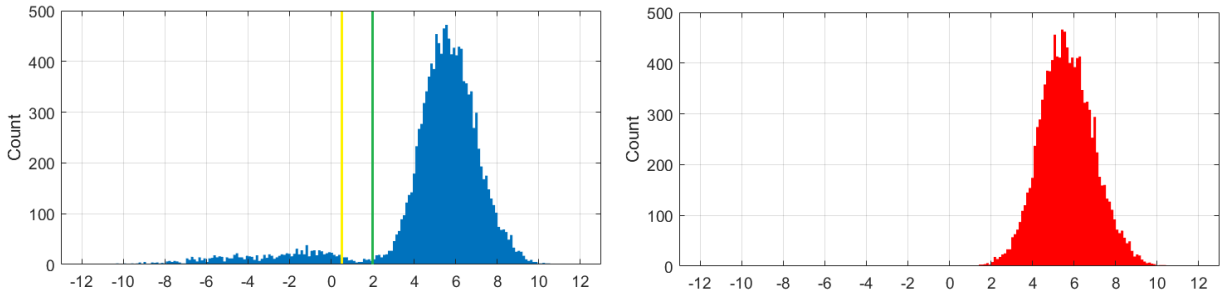

Fig 2. Distribution of biometric genuine scores used in the error detection stage for the same speaker's recordings (left). The selected thresholds are highlighted with yellow and green lines. On the right, analogous distribution after the inaccuracies were identified and removed from the corpus.

Gender labeling verification

To identify incorrect gender labels, we analyzed no-cross-gender verification results. For each speaker, a representative recording was selected and biometrically verified against representative recordings of other speakers within the same labeled gender. We then calculated the average comparison result for each speaker with their same-gender counterparts. These averages were plotted to visualize the distribution across all speakers. The distribution highlighted outliers with the lowest mean comparison values. Based on this, we set a threshold for the average log-likelihood ratio. Recordings below this threshold underwent listening analysis for gender identification. The listening results were compared with the database's gender tags. If a label inaccuracy was found, it was corrected.

All voice print comparisons in this stage were impostor scores, involving different speakers. Our methodology is grounded in the observation that cross-gender comparisons typically yield lower average results due to significant

voice characteristic differences between genders. Consequently, speakers with incorrect gender labels should contain notably lower values in no-cross-gender comparison groups.

Search for recordings of the same speakers

To find recordings of the same speakers in folders assigned to other speakers, we performed a no-cross-gender verification analysis. A representative recording was selected for each speaker and verified against recordings of other speakers within the same gender, using the speech corpus with corrected gender labels. Based on the analysis results, we plotted a data distribution. From this plot, we set a threshold value for the log-likelihood ratio. Recordings with results above this threshold were analyzed further to determine if they came from the same speaker. During listening analysis, pairs of recordings were assigned one of three labels:

- SAME SPEAKERS - indicating the recordings are from the same speaker,
- DIFFERENT SPEAKERS - indicating the recordings are from different speakers,
- NOT SURE - indicating uncertainty about the origin of the recordings.

We first addressed the pairs labeled 'NOT SURE.' Three options were considered for these recordings:

- Deleting both folders,
- Keeping both folders intact,
- Deleting one of the folders.

This method facilitates accurate classification of recordings and improves the reliability of the speech corpus.

The first option, removing both folders with problematic recordings, could artificially lower error rates in biometric software tests if recordings are from the same speaker. It might also reduce a program's ability to distinguish between similar voices when developing new solutions. The second option, keeping both folders, might artificially increase error rates if the recordings are from different speakers, leading to challenges in developing new biometric or human speech generation software. Therefore, we chose a middle ground: removing the folder with fewer recordings. This approach retains more recordings from a broader range of speakers while mitigating the issues of the first option (removing huge amount of data).

Pairs of recordings labeled 'DIFFERENT SPEAKERS' were kept separate, indicating different speakers. Those labeled 'SAME SPEAKERS' were merged into one folder, with the other folder removed. For larger groups with the 'SAME SPEAKERS' label, all associated folders were also combined into one.

This methodology primarily addresses errors where different labels are assigned to folders containing recordings from the same speaker. To reallocate misclassified recordings to correct folders, we repeat the process and move 'SAME SPEAKERS' recordings accordingly. However, due to the large number of misplaced recordings in the Clarin-PL Studio Corpus, this would significantly increase the time required for listening analysis.

Verification of the modifications performed

To partially verify the modifications from previous stages, impostor verification scores were analyzed. Each representative recording was compared against recordings in other folders. The results were sorted by the verification score, with the highest-scoring audio pairs undergoing listening analysis to ensure correct placement. This step allowed for further identification and removal of misclassified recordings.

For additional verification, the described stages could be repeated on the modified speech corpus.

Summary

This methodology outlines one approach to identify and correct inaccuracies in speech corpora, focusing on biometric recognition features. The proposed algorithm can be adapted and its stages reordered to suit different databases and modification goals. While human listening analysis is essential at each stage, the primary aim of this solution is to automate the selection of problematic recordings for review, thereby minimizing manual listening effort.

It's important to note that the Clarin-PL speech corpus, on which this methodology was applied, was recorded in a studio environment using two types of microphones, a high-quality studio mic and a standard consumer headset (Korzinek et al., 2017), This resulted in recordings of good quality with low noise. The effectiveness of this methodology might be reduced in corpora with noisy, low-quality recordings. Our tests on similar noisy corpora indicated that unequivocally categorizing recordings as from the same or different speakers can be challenging.

The success of this automated selection process relies heavily on the biometric recognition software used. For enhanced accuracy and reliability, it is advisable to use multiple voice biometric software tools. This allows for additional verification of corpus modifications and the identification of errors not detectable with a single program.

Experiments with Clarin-PL Studio Corpus and results

Applying the algorithm to the Clarin-PL speech corpus, we verified metadata crucial for biometric recognition and personalized speech synthesis. Initially, 88 of 295 folders were flagged for sub-threshold biometric scores, leading to the automatic removal of over 1,100 files from the original 13,802. Further analysis resulted in two folders being entirely removed and the corpus reduced to 12,542 files.

In subsequent stages, gender labeling inaccuracies were corrected, and folder mergers were executed to enhance data integrity. Notably, impostor comparison reductions post-modifications were significant, illustrating the effectiveness of our approach.

The corpus's final structure featured 251 speakers and 12,295 files, marking a 12% reduction in files and size to 5.28 GB. Post-modification, Phonexia software's evaluation showed improved biometric verification performance, with notably lower EER and FAR/FRR values, as detailed in Tables 1 and 2. These improvements underscore the corpus's enhanced reliability for biometric recognition software development.

Finally, it is worth mentioning that the aforementioned inaccuracies stem from the database's originally different purpose — they become relevant only when the speech corpus is used for other applications, such as biometric identification.

Table 1: EER (Equal Error Rate) values for both corpora versions depending on the different lengths of the audio samples that were verified.

	10s	8s	6s	5s	4s	3s	2s
Original Clarin-PL Studio Corpus	2.87	2.81	2.82	2.87	2.93	3.07	3.93
Modified Clarin-PL Studio Corpus	**0.33**	**0.46**	**0.59**	**0.76**	**1.05**	**0.76**	**3.37**

Table 2: FRR (False Rejection Rate) values for both Clarin-PL corpus versions and 10 seconds long recordings at fixed FAR (False Acceptance Rate) values.

FAR	10%	1%	0.1%	0.01%
Original Clarin-PL	2.30	2.98	8.16	67.66
Modified Clarin-PL	**0.01**	**0.07**	**2.11**	**29.98**

Conclusions

This paper presents a method to streamline the identification and correction of inaccuracies in speech corpora, vital for biometric recognition and human speech synthesis software development (Fig 3). It utilizes biometric verification to detect and amend common inaccuracies such as gender mislabeling, misassignment of recordings, and mislabeling of speaker origins. Our experiments show that correcting these errors can significantly lower the EER and FAR/FRR rates, enhancing the corpus's reliability for biometric software testing and development. Despite the effectiveness, the limitation of human error in auditory analysis remains. Future work may include:

- Enhancing inaccuracies identification through multiple biometric tools,
- Advancing automation in error detection,
- Improving inaccuracies detection in low-quality corpora,
- Automating biometric score threshold selection for auditory analysis.

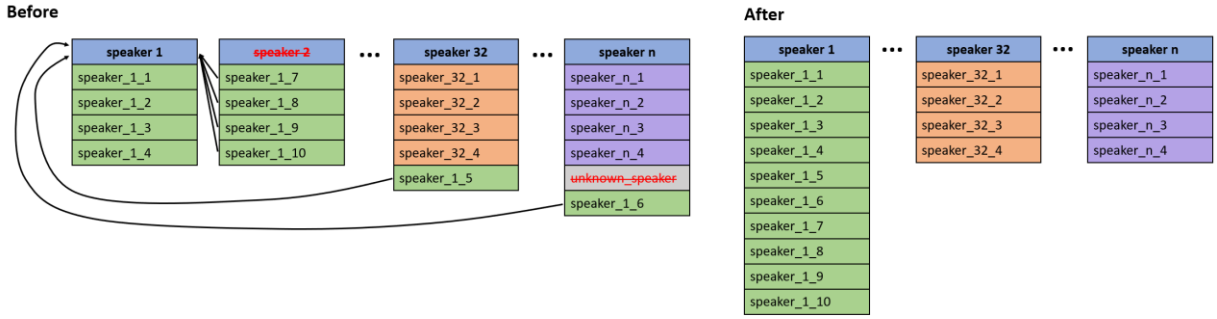

Fig 3. The left part of the figure shows the speech corpus content with e2, e3, and e4 type inaccuracies. The right part illustrates the structure of the speakers' recordings database after error elimination, including the removal of redundant directories and the reallocation of recordings to their correct folders (shown with arrows and strikethroughs in the left diagram).

Limitations

The method described in this article has certain limitations that are important to highlight and consider during its use and development.

First, it is a semi-automatic method, which means it requires manual assistance both in selecting thresholds for the biometric score (below/above which samples can be automatically discarded or need to be analyzed by listening) and in making specific decisions based on the review (gender, recordings of the same or different person). This adds

extra work that must be done during the inaccuracy search in the speech corpus. Additionally, the listening analysis itself, due to human factors, is also subject to some error and uncertainty.

Second, the accuracy of the method is limited by the accuracy of the software used to obtain the biometric comparison score of audio samples. For example, if the biometric software calculates a high similarity for audio samples from different speakers, such a case, if related to an incorrect annotation in the database, can be very difficult to identify.

References

- Budzianowski, P., Sereda, T., Cichy, T. & Vulić, I. (2024) 'Pheme: efficient and conversational speech generation,' *ArXiv*, vol. abs/2401.02839
- Campbell, J.P. (1997) 'Speaker recognition: a tutorial,' *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462. DOI: 10.1109/5.628714.
- Chakroun, R. & Frikha, M. (2023) 'A deep learning approach for text-independent speaker recognition with short utterances,' *Multimedia Tools and Applications*, vol. 82, pp. 1–23. DOI: 10.1007/s11042-023-14942-9.
- Chiari, I. (2007) 'Transcribing speech: errors in corpora and experimental settings,' *Proceedings on Corpus Linguistics*, Birmingham
- Chung, J.S., Nagrani, A. & Zisserman, A., (2018) 'VoxCeleb2: deep speaker recognition,' Proceedings of Interspeech 2018, pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
- González-Rodríguez, J., Toledano, D.T., Ortega-García, J. (2008) Voice Biometrics. In: Jain, A.K., Flynn, P., Ross, A.A. (eds) Handbook of Biometrics. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-71041-9_8
- Jessen, M., Bortlík, J., Schwarz, P. & Solewicz, Y. (2019) 'Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01),' *Speech Communication*, vol. 111. DOI: 10.1016/j.specom.2019.05.002.
- Korzinek, D., Marasek, K., Brocki, L. & Wolk, K. (2017) 'Polish Read Speech Corpus for Speech Tools and Services,' *CLARIN Annual Conference*
- Matoušek, J. & Tihelka, D. (2013) 'Annotation Errors Detection in TTS Corpora,' Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.
- National Institute of Standards and Technology (NIST), (2018), NIST 2018 Speaker Recognition Evaluation Plan, [Online], [Accessed 1 Martch 2024], https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf
- Phonexia, Phonexia Speaker Identification software website, [Online], [Accessed 1 March 2024], https://www.phonexia.com/product/speaker-identification/
- Tan, X., Qin, T., Soong, F.K. & Liu, T.Y. (2021) 'A survey on neural speech synthesis,' *ArXiv*, vol. abs/2106.15561.