TRANSFORMING AND PROCESSING THE MEASUREMENT SIGNALS

IMPACT OF AUDIO SIGNAL DURATION ON THE ACCURACY OF SPEAKER VOICE IDENTIFICATION

Volodymyr Povoroznyk, PhD student, Ihor Mykytyn, DSc, Professor
Lviv Polytechnic National University, Lviv, Ukraine
e-mail: volodymyr.b.povoroznyk@lpnu.ua, ihor.p.mykytyn@lpnu.ua

https://doi.org/10.23939/istcmtm2025.03.

Abstract. This paper investigates the capability of a system based on voice embeddings to identify speakers. We use a set of audio recordings from five speakers and construct clips of varying durations – 5 to 600 seconds. Pyannote-audio embeddings are extracted by a neural network, after which similarity coefficients are computed between embeddings of clips from the same speaker (intra-speaker similarity) and from different speakers (inter-speaker dissimilarity). We study how clip duration affects the protection zone when separating speakers into "own/other." The experiments show that there exists a certain clip duration that yields a relatively wide protection zone, which raises the probability of accurate voice-based identification. The results may be used in future research on biometric verification..

Key words: Voice biometrics, voice embedding, neural networks, speaker identification, similarity coefficient

1. Introduction

Modern voice biometric systems are increasingly used in security, authentication, and voice interfaces. Their effectiveness heavily depends on the ability to distinguish among speakers – even when only short audio clips are available. A fundamental element of such systems is the voice embedding: a vector representation of vocal characteristics that enables similarity and dissimilarity computations between voices.

A large body of research demonstrates the applicability of embeddings (e.g., x-vector, ECAPA-TDNN, pyannote-audio) to speaker identification tasks [1][2].

This paper reports results on how the width of the protection zone (safety margin) depends on the duration of speech clips used in identification. We conduct a comparative analysis of five speakers using embeddings and the following similarity (distance) measures: Bray–Curtis, Canberra, Chebyshev, Manhattan, Euclidean, cosine and correlation distances. The analysis covers both intra-speaker similarity and inter-speaker dissimilarity. Our goal is to determine the minimum clip length required for reliable speaker identification – information that is important for designing biometric systems.

2. Limitations

Among the key limitations of current speaker identification systems is a relatively high rate of falsely accepting an impostor as the genuine user (2-10%, depending on the embedding method) [3]. In addition, no publicly available datasets provide high-quality, long recordings where multiple speakers with different timbres (bass, tenor, soprano, etc.) read the exact same text – data that would be necessary for a fair comparison of protection systems that use voice identification. There is also no universal criterion for choosing the optimal similarity metric.

3. Objective

The objective is to study how the duration of audio clips affects the width of the protection zone; enlarging this zone enables improved characteristics in speaker-identification systems.

4. Research methods and parameters

We created several audio recordings and split each into equal-length clips. For every clip we computed an embedding using a pyannote-audio neural network. We then compared embeddings for clips from the same speaker (intra-speaker similarity) and clips from different speakers (inter-speaker dissimilarity). In each case, all clip embeddings from one recording were compared to all clip embeddings from a second recording (Fig. 1).

To compute similarity coefficients between two embeddings we used several standard metrics [4], specifically:

- cosine distance,
- Euclid distance,
- Manhattan distance,
- Canberra distance,
- Bray–Curtis distance,
- Chebyshev distance,
- correlation distance.

From these coefficients we derived the width of the protection zone (Fig. 2), which is directly related to "own/other" decision accuracy and stability, also we examined its dependence on clip duration.

The width of protection zone ΔZ is calculated using the formula:

$$\Delta Z = ID\_min - IS\_max \quad ,$$

where ID_min – is the minimum similarity value over the inter-speaker mismatch matrix, IS_max – is the maximum similarity value over the intra-speaker mismatch matrix.

In addition, a comparative analysis of similarity coefficients was conducted to determine the optimal metric that would provide the widest protection zone.

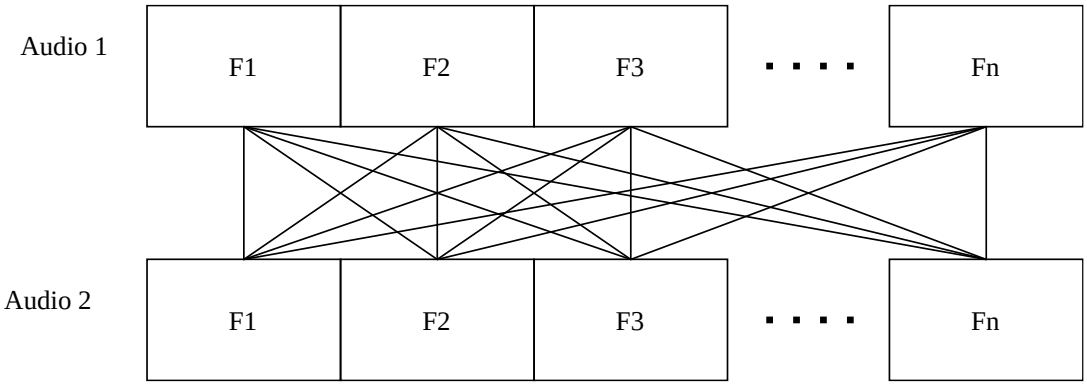


Fig1. Schematic of pairwise comparison of clip embeddings: F_i – is the i-th clip, where $i = 1,2,\ldots n$

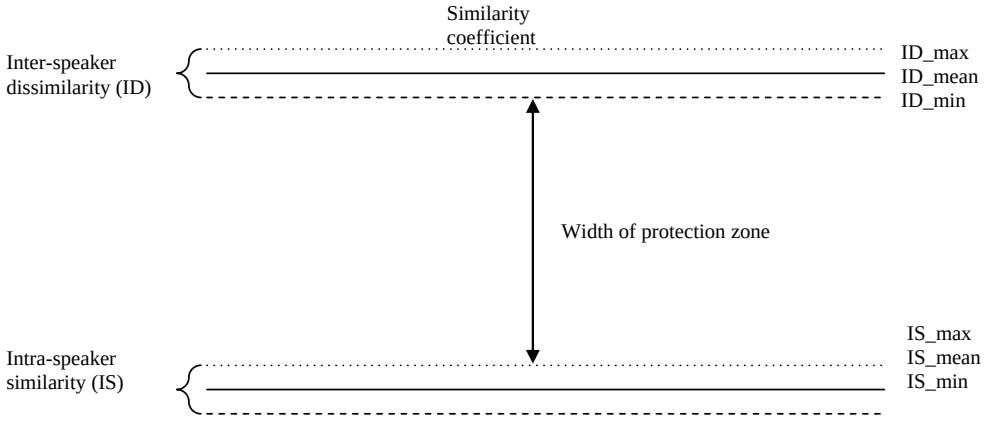


Fig. 2. Protection zone definition.

4.1. Dataset construction and embeddings calculation

The dataset comprises five audio recordings, each rendering the same English text in a different voice (two female and three male). The recordings were synthesized with Google TTS Chirp-3 [5], due to long recordings of identical text by multiple human speakers being unavailable.

Recording parameters:

- Format: WAV.
- Sample rate: 24 kHz.
- Bitrate: 384 kbit/s.
- Channels: mono.
- Duration: 60 min.

Embeddings were obtained using annote-audio [6], producing 512-dimensional vectors. Prior analysis indicated that this network performs well on long or mixed recordings [7]. We evaluated clip durations of 5, 6, 8, 10, 12, 15, 20, 24, 30, 40, 60, 144, 150, 180, 200, 225, 240, 300, 360, 400, 450, and 600 seconds. Durations were chosen so that each recording divides evenly, e.g., 720 clips of 5 s, 600 clips of 6 s, etc.

4.2. Metrics description

The **"cosine distance"** metric measures the angle between two vectors in a multidimensional space, irrespective of their lengths. It is often used in recognition tasks where the direction, rather than the vector's magnitude, is crucial. This metric is calculated by the following formula:

$$d_{cos}(A,B) = 1 - \frac{\Sigma_{i=1}^{n} \; a_i \cdot b_i}{\sqrt{\Sigma_{i=1}^{n} \; a_i{}^2} \cdot \sqrt{\Sigma_{i=1}^{n} \; b_i{}^2}}, \qquad (1)$$

where A, B – vectors, a_i, b_i – coordinates of vectors A, B accordingly.

The **"Euclidean distance"** metric is the standard function in an n-dimensional Euclidean space; its value is computed as the square root of the sum of squared differences between the corresponding coordinates of two vectors:

$$d_{eucl}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}. \quad (2)$$

The **"Manhattan distance"** metric is defined as the sum of the absolute differences between the vectors' coordinates:

$$d_{manh}(A,B) = \sum_{i=1}^{n} |a_i - b_i|. \quad (3)$$

This metric is sensitive to rotations of the coordinate system but invariant to reflections about a coordinate axis and to translations [8].

The **"Canberra distance"** metric is a weighted, normalized form of the Manhattan distance, computed as:

$$d_{canb}(A,B) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{|a_i| + |b_i|}. \quad (4)$$

This metric exhibits heightened sensitivity to small coordinate values because each term is normalized by the sum of the components' absolute values [9].

The **"Bray–Curtis distance"** metric is symmetric and normalized, and is defined by the function:

$$d_{bc}(A,B) = \frac{\sum_{i=1}^{n} |a_i - b_i|}{\sum_{i=1}^{n} (a_i + b_i)}. \quad (5)$$

The function's value lies in the interval [0,1], where 0 indicates complete identity and 1 denotes complete dissimilarity of the two vectors.

The **"Chebyshev distance"** metric is defined as the maximum absolute difference between the corresponding coordinates of the vectors:

$$d_{cheb}(A,B) = max(|a_i - b_i|). \quad (6)$$

This metric is most commonly used in tasks where the maximum coordinate-wise deviations are of primary importance.

The **"correlation distance"** metric is defined using the Pearson correlation coefficient as:

$$d_{corr}(A,B) = 1 - \frac{\sum_{i=1}^{n} (a_i - \underline{a})(b_i - \underline{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \underline{a})^2 \cdot \sum_{i=1}^{n} (b_i - \underline{b})^2}}, \quad (7)$$

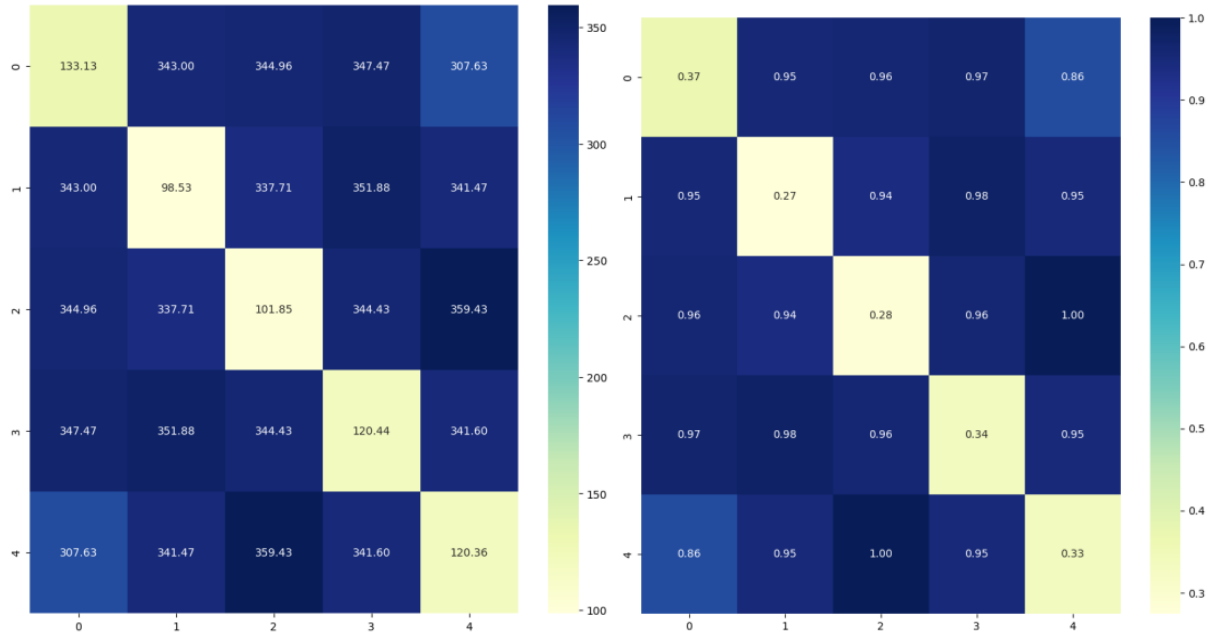
whear $\underline{a}$ та $\underline{b}$ – mean value of vector coordinates.

This metric captures only linear correlation between variables and does not account for other types of relationships.

5. Effect of audio duration on speaker identification error

Because some metrics produce values outside of [0, 1] range, we normalized all results to this range for a fair comparison (including coefficients originally in [0, 1], which rarely reach 1 exactly). The normalization procedure:

1. Construct mismatch matrices of similarity coefficients.
2. Find the maximal coefficient value – max.
3. Compute normalization factor k = 1/max.
4. Multiply all coefficients by k.



а *б*

Fig. 3. Canberra-distance mismatch matrices: (a) pre-normalization; (b) post-normalization.

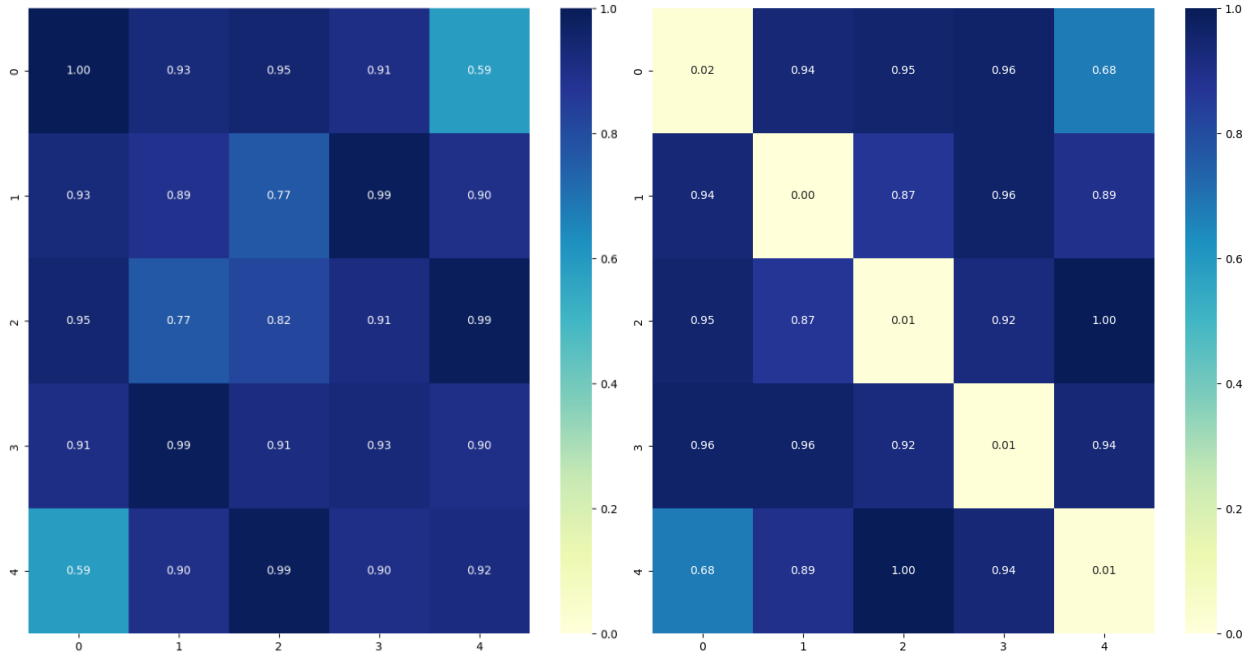
Fig. 3 illustrates Canberra-distance mismatch matrices before (a) and after (b) normalization.

Fig. 4 shows normalized cosine-distance mismatch matrices for clip durations 5 s (a) and 400 s (b).

On each matrix's main diagonal we placed the maximum intra-speaker similarity values; off-diagonal entries show the minimum inter-speaker similarity values. At 400 s (Fig. 4b) the protection zone is sufficiently wide, yielding a clear boundary between intra-speaker and inter-speaker results. At 5 s (Fig. 4a) there is practically no protection zone – some inter-speaker minima exceed intra-speaker maxima.

Table 1 reports ID_min, IS_max and ΔZ for durations 5, 20, 100, 200, 300, and 400 s across seven metrics.

Fig. 5 shows the dependence of the maximum similarity-coefficient values for intra-speaker similarity (the IS_max curve) and the minimum similarity-coefficient values for inter-speaker dissimilarity (the ID_min curve) – obtained using the cosine distance metric – on the duration of the audio segments.



а *б*

Fig. 4. Cosine-distance mismatch matrices: (a) 5 s clips; (b) 400 s clips.

Table 1. Results for ID_min, IS_max and ΔZ at selected clip durations.

	5			20			100		
Metric	ID_min	IS_max	**ΔZ**	ID_min	IS_max	**ΔZ**	ID_min	IS_max	**ΔZ**
Cosine	0,59	1	**-0,41**	0,57	0,54	**0,03**	0,55	0,16	**0,39**
Euclidean	0,77	1	**-0,23**	0,75	0,73	**0,02**	0,8	0,42	**0,38**
Manhattan	0,76	1	**-0,24**	0,74	0,72	**0,02**	0,78	0,42	**0,36**
Canberra	0,81	1	**-0,19**	0,81	0,85	**-0,04**	0,85	0,62	**0,23**
Bray-Curtis	0,68	1	**-0,32**	0,64	0,64	**0**	0,69	0,33	**0,36**
Chebyshev	0,46	1	**-0,54**	0,74	0,98	**-0,24**	0,75	0,5	**0,25**
Correlation	0,59	1	**-0,41**	0,56	0,54	**0,02**	0,63	0,18	**0,45**
	200			300			400		
Metric	ID_min	IS_max	**ΔZ**	ID_min	IS_max	**ΔZ**	ID_min	IS_max	**ΔZ**
Cosine	0,57	0,09	**0,48**	0,60	0,06	**0,54**	0,68	0,04	**0,64**
Euclidean	0,81	0,32	**0,49**	0,68	0,07	**0,61**	0,83	0,19	**0,64**
Manhattan	0,79	0,31	**0,48**	0,79	0,25	**0,54**	0,8	0,19	**0,61**
Canberra	0,85	0,52	**0,33**	0,84	0,44	**0,40**	0,86	0,37	**0,49**
Bray-Curtis	0,69	0,24	**0,45**	0,69	0,19	**0,5**	0,65	0,13	**0,52**
Chebyshev	0,71	0,33	**0,38**	0,72	0,26	**0,46**	0,7	0,19	**0,51**
Correlation	0,65	0,1	**0,55**	0,68	0,07	**0,61**	0,68	0,04	**0,64**

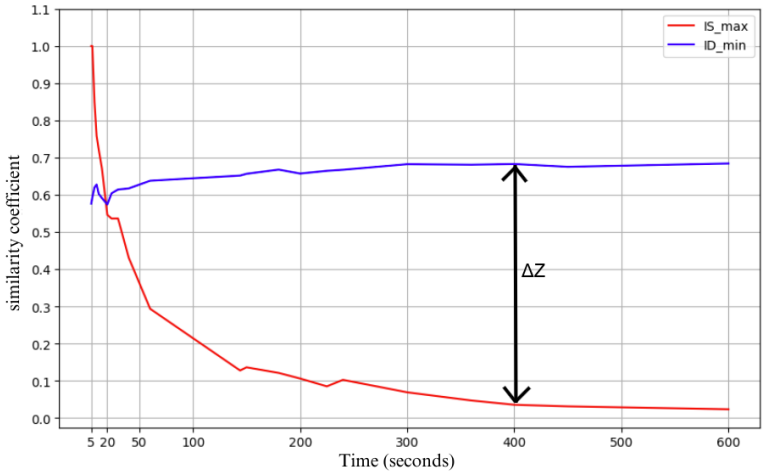


Fig. 5. Dependence of ΔZ on clip duration

As seen in Fig. 5, for 5-second clips the protection zone is negative (Table 1), indicating low identification accuracy. Increasing clip duration produces and enlarges a positive protection zone, improving the reliability of "own/other" decisions.

For durations from 5 to 20 seconds, ΔZ is negative or near zero, indicating an absent or negligible protection zone. For durations exceeding 20 seconds, a clear protection zone appears (Fig. 5). For example, with cosine distance $\Delta Z = 0{,}39$ at 100 s and $\Delta Z = 0{,}64$ at 400 s (Table 1).

Comparing metrics (Table 1), at 400 s the widest protection zones (>0,61) are achieved by four metrics: cosine, correlation, Euclidean, and Manhattan. As duration decreases, correlation distance performs best; at 100 s it yields $\Delta Z = 0{,}45$, whereas the others do not exceed 0.39.

5. Conclusion

As clip duration increases, the protection zone widens substantially, yielding higher accuracy and stability for "own/other" identification. There exists a duration (around 400 s) beyond which further increases bring only marginal gains. For 5-second clips – commonly used in speaker-ID tasks – the protection zone is negative, indicating poor accuracy. A distinct protection zone appears for durations above 20 s. For relatively long clips, cosine, correlation, Euclidean, and Manhattan distances provide the widest zones; as duration shortens, correlation distance becomes preferable. Future work will apply these findings to filter out synthesized voices during speaker identification.

Gratitude

The authors thank colleagues of the Department of Information and Measuring Technologies at Lviv Polytechnic National University for assistance in preparing the article for publication.

Conflict of Interest

The authors state that there are no financial or other potential conflicts regarding this work.

References

[1] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech 2020. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2020/pdfs/1137.pdf

[2] Bredin, H., Laurent, A., & Gillies, A. (2020). Pyannote.audio: neural building blocks for speaker diarization. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). [Online]. Available: https://ieeexplore.ieee.org/document/9053198

[3] Ruda, H., Sabodashko, D., Mykytyn, H., Shved, M., Borduliak, S., & Korshun, N. (2024). Specifics of creating and distributing phishing web resources. Cybersecurity: Education, Science, Technique, 2(12), 80–88. Available: https://csecurity.kubg.edu.ua/index.php/journal/article/view/645/508

[4] A. Levy, B. Riva Shalom, and M. Chalamish, "A guide to similarity measures and their data science applications," *Journal of Big Data*, vol. 12, Art. no. 188, 2025. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01227-1

[5] Google Cloud, "Chirp 3 HD: High-quality, low-latency text-to-speech model," *Google Cloud Text-to-Speech*, 2024. [Online]. Available: https://cloud.google.com/text-to-speech/docs/chirp3-hd

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *arXiv preprint arXiv:1911.01255*, 2019. [Online]. Available: https://arxiv.org/abs/1911.01255

[7] H. Bredin, A. Laurent, and Y. Zhong, "End-to-end domain-adversarial voice activity detection," *Proc. Interspeech 2021*, pp. 4658–4662, 2021. [Online]. Available: https://www.isca-archive.org/interspeech_2021/bredin21_interspeech.pdf

[8] E. F. Krause, *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. New York, NY, USA: Dover Publications, 1986.

[9] T. Souravlas, I. Roumeliotis, C. Roumeliotis, and C. Zissis, "Time series similarity measures and deep learning: State-of-the-art review," *arXiv preprint* arXiv:2412.20574, 2024. [Online]. Available: https://arxiv.org/abs/2412.20574