Analyzing and Improving Speaker Similarity Assessment for Speech Synthesis

Marc-André Carbonneau[1], Benjamin van Niekerk[1,2], Hugo Seuté[1], Jean-Philippe Letendre[1],
Herman Kamper[2], Julian Zaïdi[1]

[1]Ubisoft La Forge, Canada
[2]Electrical and Electronic Engineering, Stellenbosch University, South Africa

marcandre.carbonneau@gmail.com

Abstract

Modeling voice identity is challenging due to its multifaceted nature. In generative speech systems, identity is often assessed using automatic speaker verification (ASV) embeddings, designed for discrimination rather than characterizing identity. This paper investigates which aspects of a voice are captured in such representations. We find that widely used ASV embeddings focus mainly on static features like timbre and pitch range, while neglecting dynamic elements such as rhythm. We also identify confounding factors that compromise speaker similarity measurements and suggest mitigation strategies. To address these gaps, we propose U3D, a metric that evaluates speakers' dynamic rhythm patterns. This work contributes to the ongoing challenge of assessing speaker identity consistency in the context of ever-better voice cloning systems. We publicly release our code.

Index Terms: speech synthesis, computational paralinguistics

1. Introduction

Generative models that synthesize virtual humans have progressed tremendously over the last few years. Virtual characters are expected to possess unique identities that remain consistent across time. Despite recent progress, this remains a challenge in visual [1–4] and text-based systems [5, 6]. This paper considers how identity is measured in generative speech systems. Voice identity is difficult to define because it spans factors relating to anatomy, such as pitch range and timbre, as well as behavioral factors, affected by a speaker's language, accent, and natural speaking patterns and rhythm [7, 8]. Yet humans routinely identify each other under many changing conditions, implying that a set of measurable markers exists.

Characterizing identity from speech has primarily been studied for automatic speaker verification (ASV) applications [9–12], where the goal is to discriminate between speakers. This differs from modeling a speaker's identity [13]. For speaker discrimination, the models encode the *minimal* set of markers highlighting the differences between speakers, while a representation useful for generative tasks should encode *all* relevant identity markers. ASV models can therefore decide to ignore subtle speech identity factors (e.g. pitch patterns, rhythm, and accent) that are crucial for speaker reproduction.

Nevertheless, discriminative ASV representations have been shown to correlate with human judgments [14], are successfully used to model voice in few-shot speech synthesis systems [15–18], and can be used to align models [19]. In addition, these representations are routinely used to measure identity preservation of synthesis systems in research papers [16, 18–25] and community challenges [26, 27].

In this paper, we examine ASV representations through the lens of synthesis, highlighting their strengths, limitations, and implications for speech generation. We begin by reviewing identity markers in speech and identify metrics used to measure them. Next, we analyze how current ASV representations capture these markers. Our findings reveal that ASV embeddings mostly encode spectral information like pitch range and timbre, and that dynamic features are often overlooked.

We then identify confounding factors that can derail speaker similarity experiments. Specifically, we show that factors such as file duration, channel noise, and equalization can mistakenly cause speech from the same speaker to be attributed to different speakers. To address this, we provide mitigating strategies and guidelines for experimental protocols and result interpretation. While we limit our discussion to speech synthesis, our results have broader implications for speaker verification.

To address the shortcoming that ASV embeddings neglect dynamic aspects of identity, we propose U3D (Unit Duration Distribution Distance), a metric to capture speech rhythm patterns. By modeling rhythm based on phonetic content and as a distribution instead of a statistical moment, we show that U3D captures rich rhythmic information that goes beyond simple metrics like speech rate. U3D models duration over self-supervised speech units. It, therefore, does not require phonetic annotations, making it language agnostic and practical for many contexts. We open-source a framework for voice similarity measurement, implementing U3D and our prescribed best practices[1].

2. Background

2.1. Identity markers in speech

Foundational phonetic research [7, 8] identifies two primary categories of cues that enable humans to recognize each other: anatomical and behavioral factors.

Anatomical factors relate primarily to the physical structure of the vocal apparatus. The morphology, shape, and size of the vocal tract, along with muscle flexibility, largely define how a person sounds. While certain anatomical features remain relatively fixed (such as the dimensions of the nasal cavity), others—notably the oral cavity and pharynx—undergo continuous modification during speech production to modulate phonatory airflow. These physical characteristics determine the pitch range of a person and their timbre. Timbre constitutes the distinctive quality that differentiates one voice from another even when lexical content and pitch remain constant [28]. Our experiment in Section 3.2 reveals that timbre and pitch are well represented in ASV embeddings, indicating that they play a prominent role in speaker discrimination.

Behavioral factors encompass learned speech patterns and

---

[1]https://github.com/ubisoft/ubisoft-laforge-spkrid

habits developed by a speaker. While also influenced by physical properties [29], these factors relate mainly to the manner in which the vocal apparatus is used. Behavioral factors include phoneme inventory and pronunciation habits [30], natural rhythm [31,32], distinctive phonetic liaisons between words [33], individualized articulation patterns [34,35], and speech impediments. These behavioral markers reflect influences such as age [36], time-period [37], region [30,31], social factors [38,39], and individual traits, all of which play a role in shaping speaker identity. Despite extensive research on behavioral identity markers, it is still unclear how these should be modeled and measured for speech synthesis applications. In addition, these complex identity markers are challenging for human evaluators unfamiliar with the speaker [40,41], limiting the effectiveness of listening tests in assessing the voice similarity of synthesized speech. This underscores the need for automated metrics encompassing factors beyond anatomical characteristics.

2.2. Measuring speaker similarity in synthesis research

Recent years have seen growing interest in synthesis evaluation and the scrutiny of experimental protocols [42–45]. While most efforts focus on naturalness and sound quality, some propose new data and tools for similarity assessments and more [46,47].

There are different ways to report speaker similarity in synthesis papers. Some authors use predictive models to reproduce human annotations [47–49]. But most often, speaker similarity is measured by the distance between ASV embeddings, which has been shown to correlate with human judgment [14]. Some authors directly report average similarity between genuine and synthesized speech [16,17,24,25]. However, raw similarity numbers are difficult to interpret because their range varies across ASV representations and they must be put in perspective with similarities between genuine examples of the same speaker.

This is why many studies report the equal error rate (EER) of an ASV system comparing synthesized utterances with genuine utterances [14]. In practice, reporting EER translates into measuring similarity between pairs of synthesized–genuine utterances and genuine–genuine utterances. Then, a decision threshold is set to yield equal rates of false acceptances and false rejections. If synthetic utterances differ significantly from genuine ones, their similarity scores will be lower, allowing for a decision threshold that gives perfect separation and a 0% error rate. Conversely, if synthetic utterances are indistinguishable from genuine ones, it is impossible to set such a threshold, making the decision akin to flipping a coin with a 50% error rate.

3. Experiments

While ASV embeddings are widely used for speaker similarity evaluation in synthesis, very little is known about what they actually encode and how robust they are. We explore which markers are represented in some widely used ASV embeddings, and measure the effect of confounding factors. Finally, we propose a metric for rhythm that we validate through experimentation.

3.1. Models and datasets

Our goal is not to identify the best ASV technology for speech synthesis applications, but to make observations on a few representative methods. Modern representations rely on neural networks that summarize variable length utterances in fixed-length embedding vectors. The cosine similarity between these vectors indicates the predicted similarity between speakers.

We compare conventional but still widely-used ASV meth-ods such as GE2E [12] and X-Vectors [9]. GE2E uses a contrastive loss to learn discriminative embeddings, while X-Vectors are the internal representation of a classifier trained on a finite set of speakers. The more recent ResNet-TDNN [11] and ECAPA-TDNN [10] models are architectural improvements on X-Vectors, incorporating residual and skip connections and attention for pooling. We also include more recent methods relying on self-supervised learning for feature extraction because they are increasingly popular and yield state-of-the-art results [50].

We use open-source implementations from Speech-Brain[2] [51] for X-Vector, ECAPA-TDNN and ResNet-TDNN, and Resemble-ai's implementation of GE2E[3]. Finally, we use self-supervised approaches that use WavLM [52] as feature extractor: WavLM+ECAPA-TDNN, WavLM Large+ECAPA-TDNN[4] and WavLM+X-Vector[5].

We use the ARCTIC [53] and L2-ARCTIC [54] speech datasets in all our experiments. ARCTIC comprises recordings of 18 English speakers reading phonetically balanced prompts. L2-ARCTIC expands the number of speakers with 24 non-native English speakers representing a variety of accents. We chose these datasets for their diversity and recording quality. In the experiment of Section 3.2, we complement our datasets with the 1172 speakers from the LibriSpeech [55] train-clean-100 and train-clean-360 subsets, to increase diversity.

3.2. Markers captured by ASV embeddings

We start by exploring which identity markers are encoded in ASV embeddings. We do this by trying to predict handcrafted features that capture different aspects of identity from the embeddings. Table 1 summarizes the features and provides an intuition on what they measure. Anatomical characteristics, such as pitch and timbre, are captured by mean pitch and harmonic-to-noise ratio (HNR) and α-ratio. More behavioral and dynamic speech patterns are measured by speech rate, standard deviation of pitch and loudness, as well as voiced and unvoiced segment lengths. We extract speech rate using the method in [59] because it highly correlates with the ground truth syllable rate. Using REAPER[6], we compute the mean and standard deviation (std) of pitch only on voiced segments. The other features are extracted using OpenSmile [60]. Lastly, we include file duration, which should not be linked to speaker identity.

We extract ASV embeddings for each utterance in LibriSpeech, ARCTIC, and L2-ARCTIC. Taking the embeddings as input, we train a lasso regressor to predict the features for the utterance. We also experimented with a non-linear random forest regressor, which yielded results comparable to those of lasso; therefore, we report only the lasso results. We remove outliers using the inter-quartile range (IQR) method and optimized the hyper-parameters by performing 5-fold cross-validation, making sure that every speaker belongs only to a single fold. After retraining the regressor on all data, we predict the value of each handcrafted feature. Finally, we compute the coefficient of determination (r^2) between the feature values and their predictions. A perfectly predicted feature yields a score of 1.0 while uncorrelated predictions yield a score of 0.0.

Figure 1 shows that the speaker embeddings primarily encode static spectral information (mean pitch, HNR, shimmer,

---

[2] https://github.com/speechbrain/speechbrain
[3] https://github.com/resemble-ai/Resemblyzer
[4] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification
[5] https://huggingface.co/microsoft/wavlm-base-plus-sv
[6] https://github.com/google/REAPER

Table 1: *Extracted speech features and a short explanation and intuition on what they measure.*

| Markers	Description	Intuition
Duration	Length of the speech example.	This is independent from speaker identity.
Speech rate	Number of syllables per minute.	Indicates how fast a person speaks.
Voiced segment length	The average duration of a voiced segment.	Indicates how the person elongates vowel-like sounds.
Unvoiced segment length	The average duration of an unvoiced segment.	Indicates some pronunciation habits of the speaker.
Mean pitch	Average F_0 in semitones of voiced segments.	Indicates how high or low a person speaks.
Pitch std	Standard deviation of F_0.	An expressive speaker will have a high base pitch deviation.
Mean loudness	The average perceived energy in a speech signal.	Mostly indicates of the file normalization level.
Loudness std	Standard deviation of perceived energy.	Indicates if speaker is expressive.
Shimmer	Mean amplitude difference between consecutive F_0 periods.	Voice quality indicator; irregular vocal fold vibrations result in breathiness and can show poor phonation control [56].
Harmonics-to-Noise Ratio (HNR)	Energy ratio between harmonic and noise-like components.	Voice quality indicator; high HNR characterizes sonorant and harmonic voices, while low HNR denotes an asthenic voice and dysphonia [57].
α-ratio	Ratio of the energy from 50–1000 Hz and 1–5 kHz.	Reflects voice quality; bright voices have more energy in high frequencies as opposed to dark/mellow voices [58].
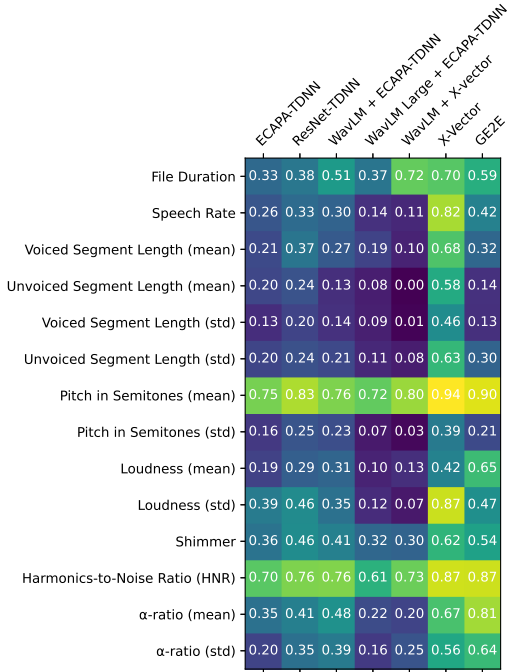
Figure 1: *The coefficient of determination (r^2) of feature predictions. A perfect score is 1.0 and uncorrelated predictions yield a score of 0.0.*

α-ratio) reflecting voice quality and frequency range. On the other hand, the embeddings fail to capture dynamic behavioral identity markers like speech rate, voiced/unvoiced segment durations, variations in pitch, and loudness. This limitation means that embeddings represent only a subset of the speaker characteristics necessary for a comprehensive assessment of identity in speech synthesis.

While our goal is not to identify the best speaker embedding to use for identity measurement, we find that X-Vector embeddings encode a wider range of markers, including more dynamic information, than more recent and better-performing embeddings for the speaker verification task. This agrees with the findings of [61]. This somewhat surprising result highlights the differences between speaker verification and speech synthesis, and calls for a more extensive comparison of ASV embeddings in the synthesis context. Better-suited identity representations might even need to be developed.

Our experiment also shows that embeddings can encode distracting factors unrelated to identity, such as speech file duration. We now explore this limitation and its implications in more detail.

3.3. Robustness of automatic speaker similarity assessment

In this section we demonstrate how ASV-based comparisons can be compromised by confounding variables, particularly file duration and channel characteristics. We do not want to invalidate established methodologies, but rather aim to guide researchers in avoiding potential methodological errors in interpreting experimental results.

Here we measure the EER (see Section 2.2) on the ARCTIC and L2-ARCTIC datasets under various conditions. Table 2 reports the average EER with the standard deviation over all speakers. As reference, we report EER after comparing each speaker to the other speakers, akin to a speaker verification task. A perfect ASV embedding would obtain an EER of 0%. As a second reference experiment, we consider each speaker individually against themselves: For each speaker, we randomly distribute speech files between two groups, the first acting as *genuine* and the second acting as *synthesized* examples. Being from the same person, both groups should be indistinguishable and EER should be 50%. Next, instead of randomly distributing the utterances, we sort them by duration, grouping shorter utterances together in one group and longer ones in another. We then measure EER for discriminating between shorter and longer utterances. Since these are the same utterances spoken by the same speaker, EER should be 50%.

As expected, control experiments with same speaker utterances yield an average EER of around 50% for all methods, and EERs between 0% and 5% when comparing different persons. This validates our protocol and shows that all embeddings accurately discriminate speakers. However, when using the same utterances but sorted by duration, EER is significantly lower for all methods. This means that these ASV systems can discriminate utterances based on file duration, which is not an indicator of identity. This coincides with the results in Figure 1, showing that ASV embeddings encode duration. This undesirable behavior can derail experiments if the duration of the synthesized examples is very different from the genuine examples: a synthesis system would under-perform, and comparisons between systems might not be meaningful. Fortunately, this can be easily mitigated with careful experimental design that ensures similar utterance durations. In the case of text-to-speech evaluation, we recommend generating the same text utterances as the genuine corpus, taking the usual precaution of not using utterances used

Table 2: *Mean and standard deviation of EER across ARCTIC and L2-ARCTIC speakers for different experiments.*

Experiment	ECAPA-TDNN	ResNet-TDNN	WavLM +ECAPA-TDNN	WavLM Large +ECAPA-TDNN	WavLM +X-Vector	X-Vector	GE2E
One speaker *vs* rest	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.03	0.03 ± 0.01	0.02 ± 0.01
Same speaker	0.50 ± 0.03	0.50 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.03
Same speaker, short *vs* long	0.34 ± 0.04	0.35 ± 0.03	0.32 ± 0.03	0.33 ± 0.03	0.39 ± 0.02	0.30 ± 0.03	0.34 ± 0.03
SNR 40	0.47 ± 0.03	0.47 ± 0.03	0.47 ± 0.03	0.48 ± 0.03	0.46 ± 0.04	0.45 ± 0.05	0.42 ± 0.04
SNR 20	0.34 ± 0.06	0.32 ± 0.05	0.33 ± 0.06	0.38 ± 0.05	0.36 ± 0.07	0.21 ± 0.07	0.15 ± 0.06
SNR 0	0.05 ± 0.04	0.04 ± 0.04	0.05 ± 0.03	0.10 ± 0.05	0.06 ± 0.04	0.01 ± 0.01	0.01 ± 0.01
+emphasis	0.47 ± 0.04	0.48 ± 0.03	0.45 ± 0.02	0.43 ± 0.03	0.48 ± 0.03	0.47 ± 0.04	0.07 ± 0.03
+de-emphasis	0.38 ± 0.05	0.41 ± 0.05	0.37 ± 0.07	0.32 ± 0.06	0.44 ± 0.06	0.28 ± 0.07	0.01 ± 0.01
+emphasis +re-equalization	0.50 ± 0.02	0.49 ± 0.02	0.49 ± 0.02	0.44 ± 0.04	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.02
+de-emphasis +re-equalization	0.50 ± 0.03	0.50 ± 0.02	0.49 ± 0.03	0.45 ± 0.05	0.50 ± 0.02	0.49 ± 0.02	0.50 ± 0.02

during training.

We next investigate channel-related confounding factors, specifically noise and equalization effects. This is important for synthesis evaluation as sound quality varies across methods. While clean, noise-free speech is desirable, quality differences should not influence speaker similarity measurements. Some systems produce artifacts that reduce sound quality compared to authentic speech, while others generate samples that are cleaner than the enrollment recordings. This latter scenario frequently occurs in voice cloning applications where users record enrollment samples with consumer-grade equipment. The 2023 Blizzard Challenge demonstrated this phenomenon, with several competitors producing samples superior to the training material [40]. To assess noise effects, we progressively introduce white noise to speech samples from each speaker and calculate EER when comparing identical utterances at varying signal-to-noise ratios (SNR). The middle section of Table 2 presents the average EER across speakers.

Similarly, we measure the effect of equalizing files by applying a popular emphasis and de-emphasis filter defined by $H(z) = 1 - \alpha z^{-1}$ with $\alpha = 0.97$ for emphasis, or its inverse for de-emphasis. These filters modify the spectral balance of the signal, without altering the identity perceived by humans. Results are given in the bottom section of Table 2.

Our experiments on channel noise (Table 2 middle section) reveal significant variations in EER following perturbation. At low SNR, all models incorrectly classify noisy samples as different identities from their clean counterparts. While 0 dB SNR represents extreme degradation, the results demonstrate that noise significantly impacts similarity scores measured by ASV systems. This has important implications when comparing systems with differing output quality, for instance when evaluating high-performance voice cloning systems using data collected in uncontrolled environments. **We recommend that researchers conduct SNR assessments on both genuine and synthesized speech** to ensure valid interpretation of the results.

Altering the spectral balance of the recording by applying equalization (Table 2 bottom section) significantly alters EER for most methods. For the GE2E implementation that we use, this even leads to catastrophic failure. Practical synthesis systems, for various reasons, can color the speech samples, making evaluation problematic. **We suggest to re-equalize the samples from the evaluated system** to get a similar spectral balance as the genuine samples. As an example, we mitigate tonal imbalance in our experiment by using a simple equalization matching algorithm. We start by estimating the power spectral densities of the genuine audio samples $S_{rr}(f)$ as well as the emphasized/de-emphasized audio samples $S_{xx}(f)$. We compute the frequency response of the corrective equalization as $|G(f)|^2 = S_{rr}(f)/S_{xx}(f)$.

Then, we obtain a 16-band FIR graphic equalizer filter [62], matching this frequency response. This filter is applied to the emphasized/de-emphasized samples to re-equalize, ensuring that it adopts the tonal characteristics of the genuine examples. In the last two rows of Table 2, we see that this brings EER back to the expected 50%.

3.4. Measuring rhythm

We have shown in Section 3.2 that ASV-based similarity evaluations only capture a subset of identity markers. They are particularly poor at measuring dynamic features. As a first step towards a better characterization of behavioral identity, we introduce U3D (Unit Duration Distribution Distance), a novel metric for modeling rhythm and phoneme duration. First, we argue that a coarse approach to measuring rhythm, based on speaking rate or mean voiced/unvoiced segment length, is insufficient for differentiating dynamic speech patterns across speakers To illustrate this, Figure 2 shows the distributions of ground-truth syllable rate and mean voiced/unvoiced segment length for the L2-ARCTIC dataset (which includes speakers from diverse origins). In many cases, adjacent speaker pairs do not show statistically significant differences in syllable rate (top). Moreover, speakers with similar syllable rates can exhibit markedly different lengths of voiced segments (bottom), suggesting that natural rhythm involves more complexity than speech rate alone can capture. These findings underscore the need for a more sophisticated evaluation method that accounts for content and is more descriptive than first or second statistical moments.

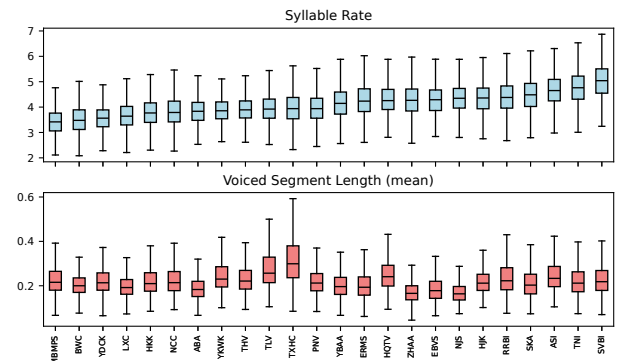We propose comparing the duration distributions of differ-

Figure 2: *Ground-truth syllable rate and voiced segment length for all speaker in L2-ARCTIC. The syllable rates are similar between many speakers, and voiced segment length (bottom) is not strongly correlated with speech rate (top).*

ent phoneme types and silence to capture finer-grained aspects of rhythm. While phoneme-duration analysis requires a forced aligner–a tool not universally available across languages–our U3D metric offers a more accessible alternative. By computing duration distributions over unsupervised speech unit groups (obtained from an SSL model), U3D provides a refined characterization of rhythm, modeling duration variations across different sound and phoneme types. This method offers researchers a robust, language-agnostic technique for quantifying the subtle variations in speakers' temporal speech patterns.

U3D is computed in three steps:

Discover phoneme-like groups of speech units: Following [59], we use agglomerative hierarchical clustering over speech units [23] extracted from a HuBERT model[7]. Different branches in the resulting dendrogram correspond with broad phoneme-like groups. For example, [59] show that the three main clusters map to sonorants, obstruents, and silences. Refer to their paper for a visualization.

Segment speech and compute duration distributions: First, we extract speech units from all speech utterances. Then, we partition them into sequences of contiguous segments using the score function from [59]. Finally, we record the segment durations to get a distribution for each phoneme-like group.

Measure distance between distributions: For each group, we compute the Wasserstein distance between the duration distributions of the synthesized and genuine speech. These distances can be reported separately for fine-grained analysis, or averaged to a give a single number for comparison.

To validate our method, we first measure duration distributions using ground-truth phoneme and silence segments obtained through forced alignment[8]. We categorize phonemes into five groups: vowels, approximants, nasals, fricatives, and stops. Then, we compute average distances across ARCTIC and L2-ARCTIC speakers under three scenarios:

1. **Same:** Compare distances between two randomly split subsets of a single speaker. This reflects the lower bound of our metric, where we expect the smallest distances.
2. **Nearest:** Calculate distance between each speaker and their closest counterpart by syllable rate. This setting tests whether our metric can distinguish between speakers undistinguishable with speech rate.
3. **Random:** Report distance between random speaker pairs.

Table 3: *Average Wasserstein distance based on forced alignments and unsupervised speech units (U3D) for the **same** speaker, between a speaker and their **nearest** neighbor by syllable rate, and between **random** pairs of speakers.*

	approx.	fric.	nasal	stop	vowel	sil.	Avg.
Forced Aligned							
Same	1.7	1.5	1.7	1.3	1.4	7.3	2.48
Nearest	8.6	8.5	7.4	7.4	7.6	70.7	18.37
Random	11.7	13.9	12.9	11.5	15.0	81.6	24.43
Unsupervised (U3D)							
Same	2.3	1.0	1.5	1.2	1.0	5.9	2.15
Nearest	11.2	4.2	5.9	12.0	5.2	71.9	18.40
Random	15.4	7.0	10.7	15.6	9.1	71.4	21.53

The results in the top section of Table 3 show that rhythm distances are significantly larger between different speakers, even

---

[7]https://github.com/bshall/urhythmic
[8]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

those with similar speech rates. Our approach therefore captures meaningful differences between speakers that cannot be reduced to speech rate. We subsequently (bottom section) replicate the experiment using unsupervised speech unit groups, confirming our method does not have to rely on forced alignment. This provides a flexible, language-agnostic technique for quantifying subtle variations in speakers' temporal speech patterns.

4. Conclusion

In this paper, we explored the limitations of ASV embeddings in the context of assessing speaker identity in speech synthesis. We showed that ASV embeddings mainly encode speech identity markers relating to anatomy (e.g. pitch range and timbre), and fail to capture time-dependent behavioral identity cues. Next, we explored the robustness of the embedding to distraction factors such as sample duration, noise level, and equalization. We provide recommendations for experimental protocol design and meaningful result interpretation. Finally, to address the shortcoming of ASV-based comparisons, we proposed U3D, a metric that better characterizes dynamic aspects of a identity by measuring rhythm. Future work includes additional metrics for behavioral markers, an extensive comparison of speaker embeddings, novel representations designed specifically for synthesis evaluation, and an extension of the discussion beyond neutral speech.

5. References

[1] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *ICLR*, 2022.

[2] N. Ruiz et al., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[3] J. Huang et al., "ConsistentID: Portrait generation with multimodal fine-grained identity preserving," *Preprint arXiv:2404.16771*, 2024.

[4] G. Wang et al., "CharacterFactory: Sampling consistent characters with gans for diffusion models," *Preprint arXiv:2404.15677*, 2024.

[5] G. Lopez Latouche, L. Marcotte, and B. Swanson, "Generating video game scripts with style," in *ACL NLP4ConvAI Workshop*, 2023.

[6] S. Han et al., "Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances," in *NACL*, 2022.

[7] P. Ladefoged and D. Broadbent, "Information conveyed by vowels." *Journal of the Acoustical Society of America*, vol. 29, 1957.

[8] K. Johnson and M. Azara, "The perception of personal identity in speech: Evidence from the perception of twins' speech," 2000.

[9] D. Snyder et al., "X-Vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020.

[11] J. Villalba et al., "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Computer Speech and Language*, vol. 60, 2020.

[12] L. Wan et al., "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[13] I. Ulgen et al., "We need variations in speech synthesis: Sub-center modelling for speaker embeddings," *Preprint arXiv:2407.04291*, 2024.

[14] R. K. Das et al., "Predictions of subjective ratings and spoofing assessments of Voice Conversion Challenge 2020 submissions," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.

[15] Y. Jia et al., "Transfer learning from speaker verification to multi-speaker text-to-speech synthesis," in *NeurIPS*, 2018.

[16] E. Cooper et al., "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP*, 2020.

[17] E. Casanova et al., "YourTTS: towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML*, 2022.

[18] C-M. Chien et al, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *ICASSP M2VoC Challenge*, 2021.

[19] S. Hussain et al., "Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance," *Preprint arXiv:2502.05236*, 2025.

[20] Q. Xie et al., "The Multi-Speaker Multi-Style Voice Cloning Challenge 2021," in *ICASSP*, 2021.

[21] W.Lin et al., "VoxGenesis: Unsupervised discovery of latent speaker manifold for speech synthesis," *Preprint arXiv:2403.00529*, 2024.

[22] S. Wang and É. Székely, "Evaluating text-to-speech synthesis from a large discrete token-based speech language model," in *LREC-COLING 2024*, 2024.

[23] B. van Niekerk et al., "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP*, 2022.

[24] M. Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale," in *NeuRIPS*, 2023.

[25] S. Chen et al., "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, 2025.

[26] W.-C. Huang et al., "The Singing Voice Conversion Challenge 2023," in *ASRU*, 2023.

[27] J.-S. Bae et al., "Generative Data Augmentation Challenge: Zero-shot speech synthesis for personalized speech enhancement," in *ICASSP GenDA Workshop*, 2025.

[28] T. F. Cleveland, "Acoustic properties of voice timbre types and their influence on voice classification," *The Journal of the Acoustical Society of America*, vol. 61, no. 6, Jun. 1977.

[29] K. Johnson, "Speech production patterns in producing linguistic contrasts are partly determined by individual differences in anatomy," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[30] M. C. Resnick, *Phonological Variants and Dialect Identification in Latin American Spanish*. De Gruyter Mouton, 1980.

[31] E. Torgersen and A. Szakay, "A study of rhythm in London: Is syllable-timing a feature of Multicultural London English?" *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, 2011.

[32] D. Deterding, "The measurement of rhythm: a comparison of Singapore and British English," *Journal of Phonetics*, vol. 29, no. 2, 2001.

[33] M. Avanzi et al., "Sociophonetics of phonotactic phenomena in french," in *International Congress of Phonetic Sciences*, 2015.

[34] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, 1993.

[35] J. S. Allen and J. L. Miller, "Individual differences in speech production: Voice-onset-time," *The Journal of the Acoustical Society of America*, vol. 108, no. 5, 2000.

[36] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *Journal of phonetics*, vol. 34, no. 4, 2006.

[37] E. R. Thomas and P. M. Carter, "Prosodic rhythm and African American english," *English World-Wide*, vol. 27, no. 3, 2006.

[38] P. Encrevé, "La liaison sans enchaînement," *Actes de la recherche en sciences sociales*, vol. 46, no. 1, 1983.

[39] J. Stuart-Smith, E. Lawson, and J. M. Scobbie, *Derhoticisation in Scottish English: A sociophonetic journey*. John Benjamins Publishing Company, 2014.

[40] O. Perrotin et al., "The Blizzard Challenge 2023," in *18th Blizzard Challenge Workshop*, 2023.

[41] ——, "Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023," *Computer Speech and Language*, vol. 90, 2025.

[42] E. Cooper, S. L. Maguer, E. Klabbers, and J. Yamagishi, "Good practices for evaluation of synthesized speech," *Preprint arxiv.2503.03250*, 2025.

[43] A. Kirkland et al., "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," in *Interspeech SSW Workshop*, 2023.

[44] P. S. Varadhan et al., "Rethinking MUSHRA: Addressing modern challenges in text-to-speech evaluation," *Preprint arXiv.2411.12719*, 2024.

[45] J. Camp, T. Kenter, L. Finkelstein, and R. Clark, "MOS vs. AB: evaluating text-to-speech systems reliably using clustered standard errors," in *Interspeech*, 2023.

[46] J. Ahn et al., "VoxSim: A perceptual voice similarity dataset," in *Interspeech*, 2024.

[47] G. Maimon, A. Roth, and Y. Adi, "Salmon: A suite for acoustic language model evaluation," in *ICASSP*, 2025.

[48] K. Deja et al., "Automatic evaluation of speaker similarity," in *Interspeech*, 2022.

[49] CH. Hu et al., "SVSNet: An end-to-end speaker voice similarity assessment model," *IEEE Signal Processing Letters*, 2022.

[50] J. Huh et al., "The VoxCeleb Speaker Recognition Challenge: a retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[51] M. R. et al., "SpeechBrain: A General-Purpose Speech Toolkit," *Preprint arXiv:2106.04624*, 2021.

[52] S. Chen et al., "WavLM: large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, 2022.

[53] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Interspeech SSW Workshop*, 2004.

[54] G. Zhao et al, "L2-ARCTIC: a non-native english speech corpus," in *Interspeech*, 2018.

[55] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[56] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Revista Brasileira de Otorrinolaringologia*, vol. 71, 2005.

[57] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, 2013.

[58] M.-J. Marsano-Cornejo and Ángel Roco-Videla, "Variation of the acoustic parameters: f0, jitter, shimmer and alpha ratio in relation with different background noise levels," *Acta Otorrinolaringologica (English Edition)*, vol. 74, no. 4, 2023.

[59] B. van Niekerk, M.-A. Carbonneau, and H. Kamper, "Rhythm modeling for voice conversion," *IEEE Signal Processing Letters*, 2023.

[60] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM-MM*, 2010.

[61] R. Fu et al., "ASRRL-TTS: agile speaker representation reinforcement learning for text-to-speech speaker adaptation," *Preprint arXiv:2407.05421*, 2024.

[62] V. Välimäki and J. D. Reiss, "All about audio equalization: Solutions and frontiers," *Applied Sciences*, vol. 6, no. 5, 2016.