The Sub-3Sec Problem: From Text-Independent to Text-Dependent Corpus

Ruichen Zuo, Kong Aik Lee, Zilong Huang, Man-Wai Mak

[1]Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University,
Hong Kong SAR, China

ruichen.zuo@connect.polyu.hk, kong-aik.lee@polyu.edu.hk, zi-long.huang@connect.polyu.hk,
man.wai.mak@polyu.edu.hk

Abstract

This paper introduces the sub-3sec problem in speaker verification, a short-duration task rarely explored. The issue arises from labor-intensive annotations and costly recordings for text-dependent speaker verification (TD-SV) corpora. To address this issue, we propose an automatic pipeline to extract short phrases from text-independent speaker verification (TI-SV) corpora. An ASR model identifies phrases and timestamps, with N-gram analysis ensuring phrases are common across speakers, enabling sufficient trials. Using this pipeline, we created Sub3Vox, a TD-SV corpus from VoxCeleb1, containing 1.6 million short utterances from 1,250 speakers—far larger than existing TD-SV corpora. Results show that matching enrollment and test phrases in TD-SV reduces EER by up to 45.23%. Additionally, shortening test utterances causes significant TI-SV performance drops but only minor reduction for TD-SV, offering the first analysis of phrase length effects on sub-3-second performance.

Index Terms: Text-dependent speaker verification, TD-SV corpus, automatic corpus curation, short-phrase speaker verification, n-gram frequency analysis

1. Introduction

Automatic speaker verification (ASV) is an identity authentication process that confirms whether a given utterance was spoken by a claimed identity [1]. It has been widely used in various real-world scenarios, including access controls, personalized services, and national security. There are two broad categories of ASV tasks [2]: text-independent speaker verification (TI-SV) and text-dependent speaker verification (TD-SV). A TI-SV system only needs to determine whether the test segment is spoken by a target speaker. The lexical contents are not taken into account in the verification process. Whereas, the content of a test utterance must match the predefined passphrase, and typically, similar to the enrollment utterance in TD-SV tasks. As such, the lexicon is restricted to a small set of predefined words or phrases in many implementations [3].

Although TI-SV is broadly studied and implemented due to its flexibility, the phonetic mismatch between enrollment and test utterances limits the performance of TI-SV systems, especially when the utterance duration is short [4]. Consequently, they are less suitable for access-control scenarios, such as digital banking, military identity verification, etc. TD-SV requires that the test utterances conform to a specific lexical content, for which test utterances of mismatched passphrases are rejected. Therefore, the context-dependent comparisons lead to higher accuracy, especially under the condition of short duration [5,6]. In recent work [7], it was reported that text-dependent verification is more reliable when facing deepfakes. These advantages make TD-SV currently the most commercially viable and popular in voice-based access control applications [8].

Similar to TI-SV, the recent advances in deep learning have inspired the research community to apply neural embedding to TD-SV. For example, the system in [9] uses ResNet-BAM as the embedding extractor and domain adversarial training to minimize the disparity between TD and TI data. The system reported in [10] employs a probabilistic linear discriminant analysis (PLDA) backend and a DenseNet front end. The method proposed in [11] learns speaker and phoneme classification simultaneously to detect impostors by identifying lexical inconsistencies. Meta-learning [12] [13], an efficient adaptation technique in low-resource scenarios, has also been used for TD-SV. For example, the three-stage pipeline in [3] enhances TD-SV performance using a tiny target-phrase dataset.

There is a lack of large corpora for TD-SV tasks. Looking back at TD corpora from the past, many of them are limited in terms of the distribution of speakers and usage scenarios [14–16] like smart living. Among them, the well-known RSR2015 [17] was recorded manually with portable devices. It involves 197,100 utterances by 300 speakers, including 157 male and 143 female speakers. Ethnic distribution of speakers mirrors Singapore's population, limiting demographic diversity. The English recordings of DeepMine [18] contains only digits and five other phrases. A crucial problem of these corpora is that they are significantly smaller than the recent TI-SV corpora, such as VoxCeleb (1,251 speakers) [19], VoxCeleb2 (6,112 speakers) [20], VoxBlink (38,000 speakers) [21], etc., because manual collection of large TD-SV corpora is time-consuming and laborious.

With the aim of overcoming the size limitation of existing TD-SV corpora, we propose an automatic pipeline to curate TD-SV corpora from TI-SV corpora. This approach addresses the challenges posed by the insufficient scale of current TD-SV corpora while reducing the need for labor-intensive manual recording. This paper introduces the Sub3Vox, a novel English corpus for TD-SV. It was generated from a TI-SV dataset by a novel automated pipeline and is larger than any existing TD-SV corpora. Notably, this is the first time that a TD-SV corpus has been created from a TI-SV corpus. We further analyze the characteristics of Sub3Vox and report its baseline performance. The proposed pipeline can be applied to other TI-SV datasets, offering a scalable solution for generating large TD-SV corpora.

Thanks to Xuerui Huang, Chong-Xin Gan, and Lishi Zuo for the help in manual check and participation in discussion.

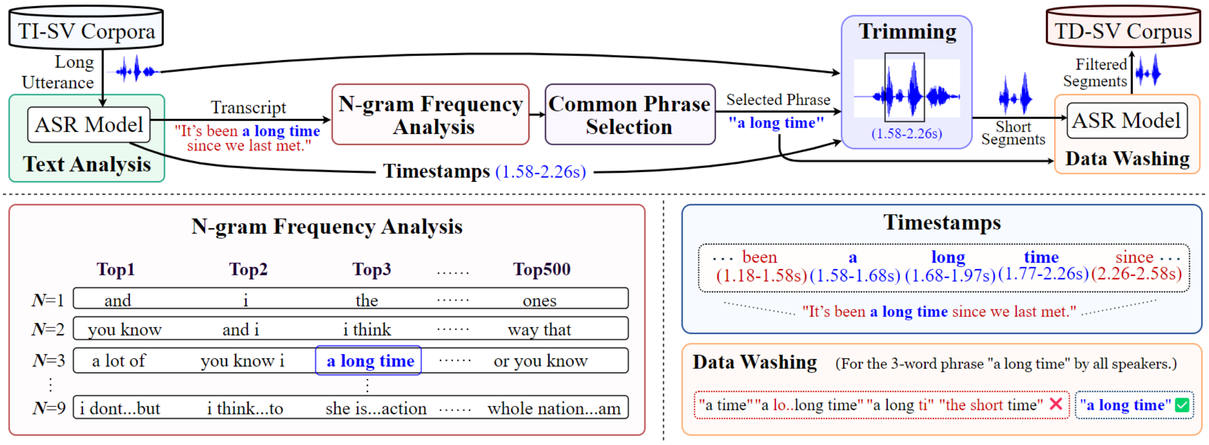
TI-SV Corpora
Long Utterance
ASR Model
Text Analysis
Transcript
"It's been a long time since we last met."
N-gram Frequency Analysis
Common Phrase Selection
Selected Phrase
"a long time"
Timestamps (1.58-2.26s)
Trimming
(1.58-2.26s)
Short Segments
TD-SV Corpus
Filtered Segments
ASR Model
Data Washing
N-gram Frequency Analysis
Top1
Top2
Top3
Top500
N=1 and i the ······ ones
N=2 you know and i i think ······ way that
N=3 a lot of you know i a long time ······ or you know
N=9 i dont...but i think...to she is...action ······ whole nation...am
Timestamps
··· been (1.18-1.58s) a (1.58-1.68s) long (1.68-1.97s) time (1.77-2.26s) since (2.26-2.58s) ···
"It's been a long time since we last met."
Data Washing (For the 3-word phrase "a long time" by all speakers.)
"a time" "a lo..long time" "a long ti" "the short time" ✗
"a long time" ✓

Figure 1: *Flowchart of the proposed automatic pipeline for creating a TD-SV corpus from TI-SV corpora.*

2. The Sub3-Second Problem in ASV

2.1. Problem Formulation

Research has shown that the speaking rate of different languages is almost the same [22]. In particular, humans speak approximately 10 phonemes per second, with an average speaking rate ranging from 3.3 to 5.9 syllables per second, depending on their emotion and other cognitive conditions [23]. As a consequence, a three-second utterance contains 9.9 to 17.7 syllables on average. Taking English as an example, the typical speaking rate in English is *four* syllables per second [24]. In this regard, we define the sub3-second problem in ASV as follows:

> **Sub3-second problem in ASV:** An ASV task with test duration of three seconds or less, without any limitation on the duration and content of the enrollment utterances. That is, the enrollment could be utterances with one or multiple utterances of the same passphrase or text-independent enrollment of long duration.

In the case of English, the Sub3-second problem requires an ASV system to make a decision based on about *twelve* or less English syllables.

2.2. Challenges in Sub3-second Verification

In the past, at least one minute of speech is required for an ASV system to achieve good performance [25]. Recent advancement in neural speaker embedding [26] has reduced the test duration to less than 10 seconds. For example, text-independent SV trained on Voxceleb2 [20] and tested on Voxeleb1 [19] can reach an EER below 1%. Today's state-of the-art ASV systems are expected to be effective when the test segments are of 3 to 10 seconds. However, test durations below 3 seconds are seldom explored in ASV. The lack of lexical coverage in short test utterances results in phonetic mismatch between the test and enrollment utterances, causing the ASV performance to degrade tremendously. On the VoxCeleb1 test set, the error rate can be increased by 368.67% when the test duration is reduced from full duration to less than 3 seconds [27].

2.3. The Importance of Sub3-second Verification

Under short-utterance scenarios, TD-SV systems generally outperform TI-SV systems, because the linguistic constraint and short utterance duration reduce the chance of phonetic mismatch [4]. Consequently, we advocate the adoption of TD-SV systems to address the sub3-sec problem.

The sub-3sec problem also motivates a new way to gather resources for TD-SV. Typically, speech corpora are collected by requiring speakers to record live through some recording devices or remotely through a telephone or mobile network [28]. Over the years, many corpora, such as TIMIT [29], RSR2015 [30], and Mixer [31], were collected in this manner and contributed significantly to advancing speaker recognition technologies. However, the laborious procedures have limited the size of these corpora. To address the sub-3sec challenges discussed in Section 2.1, it is necessary to derive a large text-dependent corpus from text-independent corpora using automated pipelines. In fact, it is fairly straightforward to extract a large number of short phrases, each with less than three seconds, that are commonly used in daily life from large TI corpora. By leveraging this method, we can effectively expand the resource base for developing TD-SV systems, thereby enhancing their performance in sub3-second scenarios.

3. Method

3.1. Deriving Large TD Corpora from TI Corpora

To overcome the size limitation of existing TD-SV corpora, we must expand speaker diversity, recording quantity, and usage scenarios. In particular, we require a substantial increase in the number of speakers and more diverse recordings. Such requirements can be fulfilled by leveraging the resources in large TI corpora.

The flowchart of our proposed automatic pipeline is shown in Figure 1. It contains four steps, which will be explained further in the following subsections.

3.2. Speech-to-Text

Text-dependent corpora put more emphasis on lexical content than text-independent ones. Extracting the text from utterances is the first step of the pipeline if the TI corpora do not have word-level transcriptions, e.g., VoxCeleb. Various self-supervised learning (SSL) front-ends can be used for this task, including wav2vec 2.0 [32], WavLM [33], HuBERT [34], and Whisper [35]. We used Whisper from OpenAI to perform speech-to-text and to obtain the timestamps of every word.

However, the Whisper model has auditory hallucination problems, which is a common issue in large speech models. Common hallucinations include transcribing the same sentence over and over again, repeating inexplicable content in non-speech regions, etc. To address this issue, voice activity detection (VAD) was used to distinguish between speech and non-speech segments in a conversation. By separating speech and non-speech segments in advance, auditory hallucinations in the Whisper model can be reduced.

3.3. N-gram Frequency Analysis

We must select as many common phrases as possible from the TI corpus to diversify the use cases of the curated TD corpus. In a TI corpus, different speakers often utter different sentences. However, in a TD system, the text of a test utterance must match the registered text of the target speaker. Therefore, it is crucial to have enough phrases spoken by the same and different speakers to form various test trials in a TD corpus. To this end, every phrase must be spoken by a speaker at least twice. We searched for the commonly used phrases, sorting the top 500 phrases for each N-word phrase, where $N = 1, 2, ..., 9$.[1]

3.4. Trimming

To obtain the recordings of the selected phrases, we trimmed the corresponding segment according to the timestamps of each phrase to form the test utterances. We made the directory structure of the proposed Sub3Vox as consistent with VoxCeleb1 as possible.

3.5. Data Washing

To reduce the detrimental effect of ASR errors on the curated dataset, we added data filtering at the end of the automated pipeline to check if the trimmed utterances correspond to the N-word phrases transcripted by the ASR model (see Figure 1). This double check procedure can further uncover some problematic and unusual cases caused by auditory hallucinations. As mentioned earlier, if the phrases were wrongly transcribed due to hallucinations, they will be different from the recognized phrases of the trimmed segments, which will be deleted by the filtering module.

4. Corpus Description

As shown in Table 1, Sub3Vox contains 1,250 speakers: 1,210 in "eval1" and 40 in "eval2", with 560 female and 690 male speakers. Compared to the original VoxCeleb, there is a slight decrease in the number of speakers because Sub3Vox includes only English utterances. We divided Sub3Vox into "eval1", and "eval2", where the utterances in Sub3Vox-eval1 were obtained from VoxCeleb1-dev and the utterances in Sub3Vox-eval2 were obtained from VoxCeleb1-test. Table 1 shows the total duration, number of unique phrases, and number of utterances in each part of the corpus. Figure 2 illustrates that most utterances in Sub3Vox are less than two seconds.

The demographic distribution in Sub3Vox is similar to that of VoxCeleb1, with most speakers from the USA and UK. Those native English speakers speak faster, making the trimmed segments shorter compared to the manual recordings, where speakers utter pre-defined phrases or sentences.

For each N from 1 to 9, we sorted the commonly used N-word phrases in the whole VoxCeleb1. Apparently, the frequency of occurrences of these N-word phrases decreases with N. Because none of the speakers in VoxCeleb1 spoke the same 9-word phrases twice, the maximum number of words in a phrase in Sub3Vox is 8. In the future, we will increase this number using a larger TI corpus, such as VoxBlink.

Table 1: *The total duration and the numbers of speakers, unique phrases, and unique utterances in each subset of Sub3Vox.*

	Male		**Female**	
	Eval1	**Eval2**	**Eval1**	**Eval2**
# of speakers	665	25	545	15
# of hours	74.48	0.51	55.29	0.33
# of unique phrases	2,302	790	2,266	622
# of unique utterances	950,680	3,457	671,798	2,108

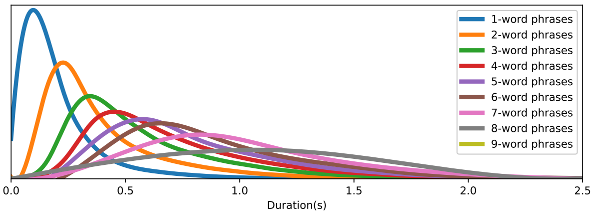

Figure 2: *Duration distributions of the phrases in Sub3Vox.*

5. Evaluation Results

5.1. Protocols of TI-SV and TD-SV

In TD-SV, both the speaker and the spoken content are considered. As shown in Table 2, TD-SV has four trial types: target-correct (TC), imposter-correct (IC), target-wrong (TW), and imposter-wrong (IW). The system accepted a test speaker only when he/she spoke the correct phrase (TC) during verification.

Table 2: *Four types of trials in TD-SV*

	Correct Passphrase	**Wrong Passphrase**
Target User	Target-Correct	Target-Wrong
Imposter	Imposter-Correct	Imposter-Wrong

5.2. Performance Metrics

The equal error rate (EER) and minimum detection cost function (minDCF) were used to measure the performance of the model. The parameter setting of minDCF follows [17], where $C_{Miss} = 10$ is the cost of missing a target speaker, $C_{FA} = 1$ is the cost of false acceptance, $P_{Target} = 0.01$ is the prior probability of target speakers, which means that the probability of the correct target speaker appearing in practical applications is 0.01. Lower equal error rate (EER) and minimum decision cost function (minDCF) indicate better performance.

5.3. Performance

In speaker verification on VoxCeleb, models are typically trained on VoxCeleb2 and tested on VoxCeleb1. Specifically, VoxCeleb2's dev and test sets are used for training, while VoxCeleb's dev and test sets are used for testing. We used pre-trained models from VoxCeleb2 in WeSpeaker [36]

[1] An N-word phrase contains a set of N words.

(ECAPA1024_LM and ResNet221_LM) to test the performance on the curated Sub3Vox. The supported scoring back-end is cosine similarity with score normalization [37].

We tested text-dependent and text-independent speaker verification to compare metrics under the same and different phonetic contexts, respectively. For enrollment, we used three different utterances of the same pass-phrase from the same speaker, with one segment as the test utterance.

5.3.1. Overall Performance

The number of trials are shown in Table 3. Results shown in Table 4 exclude 1-word phrases (such as "and", "the", etc.) because they are rarely used in speaker verification systems. Since Sub3Vox was derived from VoxCeleb1 and the models were pre-trained on VoxCeleb2, Sub3Vox can simulate real-life scenarios with unseen speakers and passwords. Testing on future Sub3Vox versions derived from VoxCeleb2 should yield better performance.

Compared with text-independent speaker verification, the performance improvement of the text-dependent method is over 30%, which is consistent across gender and evaluation subsets. The highlighted example in Table 4 reaches a performance improvement up to 45.23%, which was achieved by ResNet221-LM in the female speakers. Among the male speakers, there is also a performance improvement up to 41.54%. These results demonstrate that TD-SV has a significant advantage over TI-SV, especially when the test utterances are short.

Figure 3 shows the impact of the number of words in a phrase on the performance under TI-SV and TD-SV settings. The average utterance durations in Sub3Vox are 564ms for males and 587ms for females. At these mean durations, the expected numbers of words are 2.4 for both gender. Because the enrollment and test utterances have an integral number of words, we report the performance of TI-SV and TD-SV under 1–8 words, 2–8 words, and 3–8 words in Figure 3. This arrangement means that the evaluations on 2–8 words will exclude all 1-word phrases. Similarly, the evaluations on 3–8 words will exclude all 1-word and 2-word phrases. The results show that the TI-SV suffers from a more severe performance drop (increase in EER) when the evaluations include 1-word and 2-word phrases. The performance drop in TI-SV is even more severe when the evaluations include 1-word phrases, especially for the ResNet-211. The trend clearly suggests that TD-SV is a better choice for short-utterance scenarios.

Table 3: *The numbers of trials in the four trial types in Sub3Vox. TC: Target-correct; TW: Target-wrong; IC: Imposter-correct; IW: Imposter-wrong.*

Trial Type	**Male**		**Female**	
	Eval1	**Eval2**	**Eval1**	**Eval2**
TC	1,781,018	2,263	352,009	1,491
TW	4,096,332	3,081,496	1,508,569	2,090,158
IC	8,248,687	1,262,561	7,352,431	355,357
IW	67,264,238	40,487,487	21,031,044	17,300,245

5.3.2. Performance on Fixed Number of Words

We also conducted experiments in which all trials used utterances with a fixed number of words, e.g., 1-word phrases, 2-word phrases, and 3-word phrases. The results are shown in Figure 4. Evidently, the EER decreases when the number of words in the phrases increases. Again, the performance of TD-SV is always better than that of TI-SV.

Table 4: *EER and minDCF achieved by ECAPA-TDNN and ResNet-221 on Sub3Vox.*

Model	**Gender**	**Verf. Mode**	**Sub3Vox-eval1**		**Sub3Vox-eval2**	
			EER(%)	**minDCF**	**EER(%)**	**minDCF**
ECAPA -TDNN	Male	TI	16.90	0.66	19.05	0.74
		TD	12.25	0.52	15.07	0.62
	Female	TI	18.56	0.73	19.42	0.79
		TD	14.02	0.64	14.22	0.59
ResNet -221	Male	TI	13.00	0.58	15.05	0.57
		TD	7.60	0.26	9.55	0.40
	Female	TI	14.75	0.60	**14.90**	0.61
		TD	10.62	0.46	**8.16**	0.37

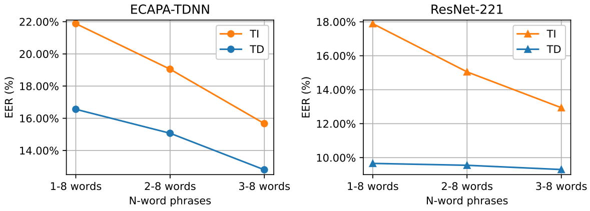

Figure 3: *The impact of phrase durations on the performance of TI-SV and TD-SV systems. In the horizontal axis, from left to right, the short phrases (1-word and 2-word phrases) are progressively excluded, leading to longer durations for the test phrases.*

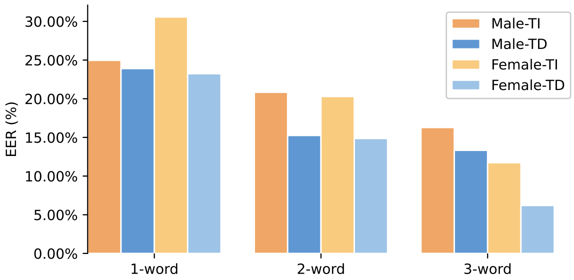

Figure 4: *The EER performance of ECAPA1024_LM on Sub3Vox-eval2 for different numbers of words in the test utterances (phrases).*

6. Conclusions and Future Works

We focus on the short-duration problems on speaker verification, and defined a new concept – the sub-3sec problem. An automatic pipeline was proposed to curate TD-SV corpora from TI-SV corpora. With this pipeline, we introduced a new dataset – Sub3Vox, derived from VoxCeleb1, and reported baselines on it. We found that the EER of a text-independent speaker verification system can be reduced by up to 45.23% when it was modified to a text-dependent one, where lexical content can play an important role in the verification process. Moreover, we show that with test utterances under 3 seconds, the EER of a TD-SV system can be reduced by up to 53.56% when the number of words per utterance was increased by one.

Future research will focus on leveraging the proposed pipeline to create more extensive TD-SV corpora from larger

TI-SV datasets, such as VoxBlink and SRE. This effort aims to further advance the development of short-duration text-dependent speaker verification systems.

7. References

[1] K. A. Lee, O. Sadjadi, H. Li, and D. Reynolds, "Two decades into speaker recognition evaluation - are we there yet?" *Computer Speech & Language*, vol. 61, p. 101058, 2020.

[2] M.-W. Mak and J.-T. Chien, *Machine Learning for Speaker Recognition*. Cambridge University Press, 2020.

[3] W. Lin and M.-W. Mak, "Model-agnostic meta-learning for fast text-dependent speaker embedding adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1866–1876, 2023.

[4] M. Hébert, "Text-dependent speaker recognition," *Springer Handbook of Speech Processing*, pp. 743–762, 2008.

[5] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SDSV) challenge 2021: the challenge evaluation plan," *arXiv preprint arXiv:1912.06311*, 2019.

[6] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[7] A. Firc and K. Malinka, "The dawn of a text-dependent society: deepfakes as a threat to speech verification systems," in *Proceedings of the 37th ACM/SIGAPP symposium on applied computing*, 2022, pp. 1646–1655.

[8] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. IV–4072.

[9] L. Zhang, J. Wu, and L. Xie, "NPU speaker verification system for Interspeech 2020 far-field speaker verification challenge," *arXiv preprint arXiv:2008.03521*, 2020.

[10] Z. Chen and Y. Lin, "Improving x-vector and PLDA for text-dependent speaker verification." in *Proc. Annual Conference of the International Speech Communication Association*, 2020, pp. 726–730.

[11] Y. Liu, Z. Li, L. Li, and Q. Hong, "Phoneme-aware and channel-wise attentive learning for text dependent speaker verification," *arXiv preprint arXiv:2106.13514*, 2021.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[13] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[14] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "The realistic multi-modal valid database and visual speaker identification comparison experiments," in *Proc. 5th International Conference on Audio-and Video-Based Biometric Person Authentication*, 2005.

[15] R. H. Woo, A. Park, and T. J. Hazen, "The MIT mobile device speaker verification corpus: Data collection and preliminary experiments," in *Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[16] J. Fierrez, J. Ortega-Garcia, D. T. Toledano, and J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database," *Pattern Recognition*, vol. 40, no. 4, pp. 1389–1392, 2007.

[17] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases," in *Proc. Annual Conference of the International Speech Communication Association*, 2012, pp. 1578–1581.

[18] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in persian and english." in *Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Annual Conference of the International Speech Communication Association*, 2017, pp. 2616–2620.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Proc. Annual Conference of the International Speech Communication Association*, 2018, pp. 1086–1090.

[21] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, and M. Li, “Voxblink: A large scale speaker verification dataset on camera,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10 271–10 275.

[22] S. Kowal, R. Wiese, and D. C. O’Connell, “The use of time in storytelling,” *Language and Speech*, vol. 26, no. 4, pp. 377–392, 1983.

[23] S. Arnfield, P. Roach, J. Setter, P. Greasley, and D. Horton, “Emotional stress and speech tempo variation,” *Speech under Stress*, pp. 13–15, 1995.

[24] A. Cruttenden, *Gimson’s Pronunciation of English*. Routledge, 2014.

[25] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “I-vector based speaker recognition on short utterances.” in *Proc. Annual Conference of the International Speech Communication Association*, 2011, pp. 2341–2344.

[26] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[27] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva *et al.*, “Deep speaker embeddings for far-field speaker recognition on short utterances,” *arXiv preprint arXiv:2002.06033*, 2020.

[28] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. Van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, “The RedDots data collection for speaker recognition,” in *Proc. Annual Conference of the International Speech Communication Association*, 2015, pp. 2996–3000.

[29] V. Zue, S. Seneff, and J. Glass, “Speech database development at mit: Timit and beyond,” *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[30] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR 2015,” *Speech Communication*, vol. 60, pp. 56–77, 2014.

[31] C. Cieri, L. Corson, D. Graff, and K. Walker, “Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora.” in *Annual Conference of the International Speech Communication Association*, 2007, pp. 950–953.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[33] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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, pp. 1505–1518, 2022.

[34] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[36] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, “WeSpeaker: A research and production oriented speaker embedding learning toolkit,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[37] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, “Analysis of score normalization in multilingual speaker recognition.” in *Annual Conference of the International Speech Communication Association*, 2017, pp. 1567–1571.