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

Ruichen Zuo, Kong Aik Lee*, Man-Wai Mak*, Zilong Huang
Dept. of Electrical and Electronic Engineering
The Hong Kong Polytechnic University
Hong Kong SAR, China
{ruichen.zuo, zi-long.huang}@connect.polyu.hk, {kong-aik.lee, man.wai.mak}@polyu.edu.hk

***Abstract*— This paper first defines the short-duration speaker verification problem, the sub-3sec problem, which has rarely been discussed in previous speaker recognition research. Then, we propose a fully automatic pipeline to derive a text-dependent speaker verification corpus from a text-independent speaker verification corpus. Using this approach, we curate the Sub3Vox corpus from VoxCeleb1. It consists of 1,250 speakers, 250.43 hours of speech, and 1,769,131 unique utterances. The number of speakers and the number of unique utterances are nearly the same as in the original VoxCeleb1, while the total speaking hours are less because they are short utterances trimmed from VoxCeleb1. We report the baselines for speaker verification tested on the proposed Sub3Vox corpus.**

***Index Terms*—Database, Corpus, Speaker Verification, Speaker Recognition, Text-Dependent**

I. Introduction

Automatic speaker verification (ASV) is a biometric authentication process of confirming whether a given utterance was spoken by a claimed identity [1] [2]. 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 [3]: text-independent speaker verification (TI-SV) and text-dependent speaker verification (TD-SV). In the former, the TI-SV system only needs to determine whether the test segment is spoken by the target speaker. The lexical contents of the enrollment and test utterances do not need to match. Whereas, the content of the enrollment and test utterances must match in TD-SV task. As such, the lexicon is restricted to a small set of predefined words or phrases in many implementations [4].

Although TI-SV is broadly studied and implemented due to its flexibility, the phonetic mismatch between enrollment and test segments limits the performance of TI-SV systems, especially when the utterance duration is short [5]. Consequently, they are less suitable for access-control scenarios, such as bank locks, military identity verification, etc. TD-SV requires that the test segments conform to a specific phonetic context, for which test utterances of mismatched passphrases are rejected. The context-dependent likelihood test leads to higher accuracy, especially under the condition of short duration [6] [7]. In a recent work [8], 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 [9].

Recent advances in TD-SV algorithms have achieved good performance. The system in [10] uses ResNet-BAM as the embedding extractor and domain adversarial training to minimize disparity between TD and TI data. The system reported in [11] employs a probabilistic linear discriminant analysis (PLDA) model as the backend and a DenseNet as the frontend. [12] introduced a method that learns speaker and phoneme classification simultaneously to detect impostors by identifying lexical inconsistencies. Meta-learning, an efficient adaptation technique in low-resource scenarios, has also been used in TD-SV systems. For example, the information-theoretic meta-regularizer [13] addresses the issue of meta-learners memorizing meta-training tasks but struggling to adapt to new tasks. The three-stage pipeline model in [4] enhanced TD-SV performance with tiny target-phrase datasets.

Whereas, there is a lack of large corpora for TD-SV tasks. Looking back at TD corpora from the past, many of them are for specific uses with a biased part of speakers and utterances [14]–[16]. Among them, the speech in the relatively large and well-known RSR2015 [17] was recorded manually by portable devices. It involves 300 speakers, including 157 men and 143 women. The ethnic distribution of speakers is close to the distribution of Singapore's population, resulting a certain degree of limitation in geographical distribution. The DeepMine [18] contains 370,000 recordings of English and Persian for both text-dependent and text-independent use. The crucial problem is that they are not as large as the corpora for TI-SV tasks, such as VoxCeleb [19], VoxCeleb2 [20], VoxBlink [21], etc., because manual collection of large TD-SV corpora is time-consuming and laborious. In this paper, we propose an automatic pipeline to curate TD-SV corpora from TI-SV corpora, which addresses the shortcomings of the existing TD corpora not being large enough and saves the effort in manual recording.

The novelties of the paper are highlighted as follows.

- The sub-3sec problem is formally defined as a speaker verification task with test segments of three seconds or less, without any limitation on the duration and context of the enrollment utterances. That is, the enrollment could be utterances with one or multiple utterances of the

* The corresponding authors are Kong Aik Lee and Man-Wai Mak.
* Thanks to Xuerui Huang, Chong-Xin Gan, and Lishi Zuo for the help in manual check and participation in discussion.

same passphrase or text-independent enrollment of long duration.

- The Sub3Vox, a new English corpus, is introduced for TD-SV. It was created from a TI-SV dataset in an automated pipeline and is larger than any existing TD-SV corpora. This is the first time that a TD-SV corpus has been created from a TI-SV corpus. The characteristics of the Sub3Vox are analyzed, and its baseline performance is reported.

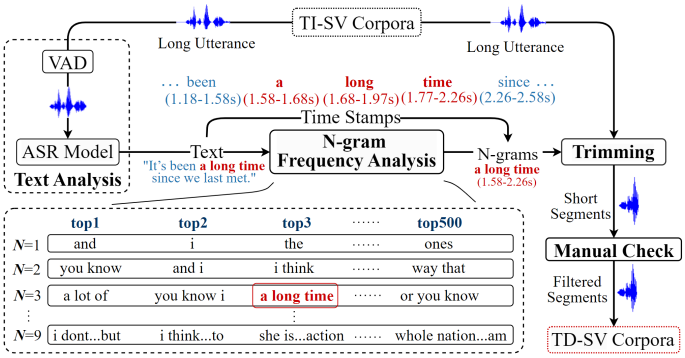

Fig. 1. Flowchart of the proposed automatic pipeline for creating a TD-SV corpus from a TI-SV corpus.

II. The Sub-3Sec Problem

In the past, at least one minute of speech is required for a TI-SV system to achieve good performance [22]. Today's state-of-the-art TI-SV systems are proven to be effective when speech segments have three to ten seconds, primarily because of sufficient coverage of the phonetic space for such a long segment. However, segments below three seconds are seldom explored by text-independent methods. The lack of lexical coverage in short utterances results in phonetic mismatch between the enrollment and test utterances, causing the performance of TI systems degrades tremendously. Nevertheless, under the same short-utterance scenario, TD-SV systems tend to perform better than TI-SV systems because the linguistic constraint and short utterance duration reduce the chance of having phonetic mismatch [5].

Although TD-SV systems have more use cases and higher commercial value than TI-SV systems, the resources for training TD-SV systems are either proprietary or very limited. To leverage the advantage of TD-SV, we advocate the sub-3sec problem that will facilitate the curation of more resources for TD-SV and reflect the practical issues of TD-SV under real-world scenarios. The problem centers around the duration for which TD-SV is viable, which is based on the following analysis of spoken languages. Research has shown that the speaking rate of different languages is almost the same [23]. In particular, humans speak approximately 10 phonemes per second, with a rate ranging from 3.3−5.9 syllables per second, depending on the condition [24]. As a consequence, three seconds of utterance contain 9.9−17.7 syllables. Taking English as an example, the typical speaking rate in English is 4 syllables per second [25]. Therefore, in the case of English, the sub-3sec problem requires a speaker verification system to make a decision based on about 12 or less English syllables.

III. Method

A. Derive TD Corpora from TI Corpora

The sub-3sec problem motivates a new way to gather resources for text-dependent verification. Typically, speech corpora are curated by requesting speakers to speak live through some recording front-ends or remotely through a telephone or mobile network [26]. Over the years, many corpora, such as TIMIT [27], RSR2015 [28], and Mixer [29], have played a critical role in advancing speaker recognition technology. However, the laborious manual work always limit the size of these corpora compared with the automatically curated ones, such as the VoxCeleb [19]. To solve the sub-3sec problem mentioned in Section II, it is possible to derive a large text-dependent corpus from text-independent corpora using automated pipelines, since it is easy to find enough short phrases less than three seconds that many people would say in daily life from large TI corpora. On the other hand, number of longer phrases (such as "open the refrigerator to get some vegetables and drinks") is much smaller.

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.

B. Text Analysis

Text-dependent corpora focus more on lexical content than text-independent ones. To develop a TD corpus, the first step is selecting available passphrases. If the TI corpus lacks transcriptions, such as VoxCeleb, transcribing the text from utterances is necessary. Various self-supervised learning (SSL) front-ends can be used for automatic speech recognition (ASR) tasks, including wav2vec 2.0 [30], WavLM [31], HuBERT [32], and Whisper [33]. We used Whisper from OpenAI for ASR on text-independent corpora.

However, hallucinations are a common issue in large models, including the Whisper model, which often repeats sentences or transcribes non-existent content. To address this issue, voice activity detection (VAD) is used to distinguish between speech and non-speech segments in a signal. By separating these parts in advance, VAD reduces auditory hallucinations in the Whisper model and improves recognition speed.

C. N-gram Frequency Analysis

After obtaining the text content from utterances in the TI corpus, we selected suitable phrases as mentioned in Section III-B. In a TI corpus, different speakers often utter different sentences, but for a TD-SV 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 text-dependent corpus. We searched the most commonly used phrases, sorting the top 500 in each N-gram list ($N = 1, 2, ..., 9$). N-gram refers to a set of N words. For text-dependent use, phrases must be spoken by a speaker at least twice.

D. Trimming

With the most commonly used phrases among the corpus, we trimmed the corresponding segment according to the timestamps of each phrase to form the test utterances. When processing the data from VoxCeleb1, we make the folder storage structure of the proposed Sub3Vox as consistent as possible with the original corpus.

E. Manual Check

To ensure the correctness of the dataset, we added a manual check at the end of the automated pipeline. This procedure can detect some problematic and unusual cases in the TI-SV corpus. For example, in VoxCeleb1, the folder of a female speaker (id10384) contains a male speaker's clip, which our pipeline could not automatically detect. Therefore, random checking by human listeners is necessary to minimize the number of incorrectly annotated segments.

IV. Corpus Description

A. Data Overview

As shown in Table I, Sub3Vox contains 1,250 speakers: 1,210 in the dev set and 40 in the test set, with 560 female and 690 male speakers. Compared to the original VoxCeleb2, there is a slight decrease in the number of speakers because Sub3Vox includes only English utterances, excluding speakers with samples in other languages. Table I also 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 one second.

TABLE I
THE NUMBER OF SPEAKERS, TOTAL HOURS, UNIQUE PHRASES, AND UNIQUE UTTERANCES IN EACH SUBSET OF SUB3VOX.

	Male		Female	
	dev	**test**	**dev**	**test**
# of speakers	665	25	545	15
# of hours	136.05	4.75	106.79	2.84
# of unique phrases	2,585	1,529	2,583	1,504
# of unique utterances	969,070	34,786	745,694	19,581

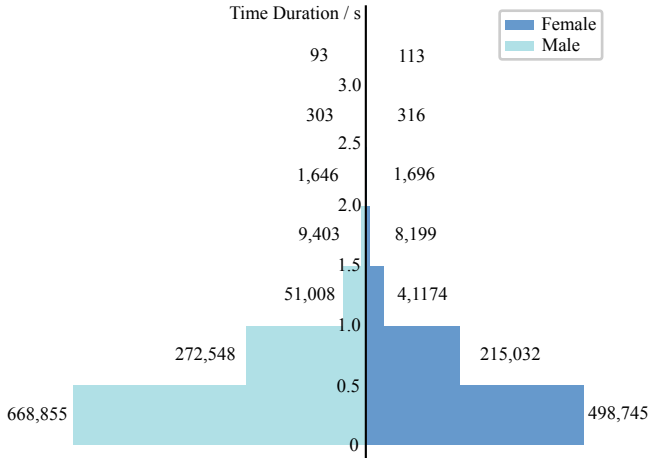

Fig. 2. Duration distribution of the utterances in Sub3Vox.

The population distribution in Sub3Vox is similar to VoxCeleb1, with most speakers from the USA and UK. As native English speakers, they speak faster and have more co-articulation, making the trimmed segments more ambiguous compared to manual recordings where speakers utter specific phrases or sentences.

We sorted commonly used N-grams in each part of the corpus, as shown in Table II. While we searched for phrases with N ranging from one to nine, higher N-grams naturally appear less frequently. No 9-gram phrases from VoxCeleb1 are spoken by the same speaker at least twice, though they may exist in larger corpora like VoxCeleb2. Additionally, 1-gram phrases (such as "and", "the") are too short and vague, due to the occurrence in consecutive sentences and co-articulation. Therefore, they are excluded from the speaker verification trials.

TABLE II
THE TOTAL DURATION (IN SECONDS) IN EACH KIND OF N-GRAM PHRASES.

N-gram	**Male**		**Female**	
	dev	**test**	**dev**	**test**
1-gram	193,030.88	7,291.32	150,854.48	4,176.89
2-gram	180,824.57	6,701.75	142,320.39	3,988.76
3-gram	78,075.51	2,633.54	61,241.77	1,669.03
4-gram	27,481.45	397.98	21,764.8	305.60
5-gram	8,485.60	57.10	6,859.87	59.60
6-gram	1,670.12	14.66	1,239.66	17.52
7-gram	190.60	3.84	157.28	1.88
8-gram	15.64	0	11.14	0
9-gram	0	0	0	0

B. Types of Errors and the Others

In addition to situations originally exist in VoxCeleb, such as background noise and mixing of voices from other speakers, we classified the recordings in the new corpus into six kinds to ensure robustness and real-world applicability. They are pre-ambiguous, post-ambiguous, pre- and post- ambiguous, stammer, clear, and wrong. As described in Section III-E, two independent checkers sampled the dev set of Sub3Vox to assess the trimmed segments. Two random and independent checks are detailed in Fig. 3, with Checker 1 (left) sampling 538 utterances and Checker 2 (right) sampling 736. Clear samples precisely match the phrases without stammering or co-articulation, while wrong samples have completely different pronunciations from their label phrases and should be filtered out.

Most wrong samples result from fast and ambiguous pronunciation, with only a few caused by hallucinations from the speech recognition model. For instance, among 735 files checked by Checker 2, only one has a completely different label due to hallucination. The other nine wrong samples have similar phrases to the target utterances, such as "I ought to" labeled as "I had to" and "there" labeled as "they were," making up just 0.14% of the total samples checked. The other four types of errors are described as follows.

1) Pre-ambiguous: Pre-ambi refers to ambiguous pronunciation at the beginning of utterances, such as an extra "for" before the 3-gram "a long time", or incomplete pronunciation at the beginning, like missing "a" before the tri-gram "a long time."

2) Post-ambiguous: Post-ambi refers to ambiguous pronunciation at the end of utterances, such as the 3-gram "a long

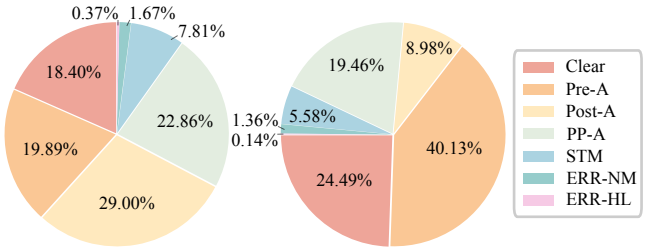

Fig. 3. Distribution of the types of errors and others to be identified by the two checkers in the dev set of Sub3Vox. Pre-A is pre-ambiguous, Post-A is post-ambiguous, PP-A is pre- and post-ambiguous, STM is stammer, ERR-NM is wrong sample by similar utterances, ERR-HL is wrong sample by hallucination.

time" followed by an extra "we", or incomplete pronunciation at the end, like missing the latter half of "time" in "a long time."

3) Pre&Post-ambiguous: Post&Post-ambi is ambiguous pronunciation at the beginning and the end of utterances simultaneously, which means having both situations mentioned above.

4) Stammer: Stammer is easy to understand. For example, the tri-gram "a long time" might be pronounced as "a, a long time." Some stammer samples are included in the corpus for robustness.

V. Evaluation Results

A. Protocols of TI-SV and TD-SV

In TI-SV, test results are either target or non-target. In TD-SV, both the speaker and the spoken content are considered. As shown in Table III, TD-SV has four trial types: target-correct (TC), imposter-correct (IC), target-wrong (TW), and imposter-wrong (IW). The system only passes when the target speaker says the registered phrase (TC).

TABLE III
4 TYPES OF TRIALS IN TD-SV.

	Correct Pass-phrase	Wrong Pass-Phrase
Target User	Target-Correct	Target-Wrong
Imposter	Imposter-True	Imposter-Wrong

B. Performance Metrics

The equal error rate (EER) and minimum detection cost function (minDCF) are used to measure the performance of the model. The detection cost function C_{det} is shown as 1.

$$C_{det} = C_{Miss}P(Miss|Target)P_{Target} + C_{FA}P(FA|NonTarget)(1-P_{Target}) \quad (1)$$

where $C_{Miss} = 10$ is the cost of missing the real target speaker, $C_{FA} = 1$ is the cost of false acceptance of the imposter, $P_{Target} = 0.01$ is the prior probability of the target, which means that the probability of the correct target speaker appearing in practical applications is 0.01.

C. 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 [34] to test performance on curated VoxCeleb1 corpora, specifically ECAPA-TDNN1024-LM and ResNet221-LM. The supported scoring back-end is cosine similarity with score normalization [35].

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. Each trial consists of enrollment and test. The number of trials, shown in Table IV, includes roughly ten times more imposter trials than target trials, similar to the RSR2015 dataset. Results are shown in Table V. Lower equal error rate (EER) and minimum decision cost function (minDCF) indicate better performance. Since Sub3Vox is derived from VoxCeleb1 and the model is pre-trained on VoxCeleb2, it simulates real-life scenarios with unseen speakers and passwords. Testing on future Sub3Vox versions derived from VoxCeleb2 should yield better performance with lower EER and minDCF.

TABLE IV
TRIAL NUMBERS OF THE FOUR LABELS IN TD-SV.

Trial Types	Male		Female	
	dev	test	dev	test
TC	20,582	21,569	21,340	13,172
TW	813,767	823,040	811502	638,479
IC	148,134	74,726	72,788	61,805
IW	9,093,204	7,997,877	6,273,609	5,613,839

TABLE V
PERFORMANCE OF EER% AND MINDCF IN EACH PART OF THE SUB3VOX. ECAPA IS ECAPA-TDNN1024-LM, RESNET IS RESNET221-LM.

Model	TD/TI	Metric	Male		Female	
			dev	test	dev	test
ECAPA	TD	EER	12.25	14.71	12.48	12.96
		minDCF	0.52	0.52	0.51	0.48
	TI	EER	16.90	19.30	15.98	14.15
		minDCF	0.66	0.75	0.64	0.60
ResNet	TD	EER	8.30	11.88	7.96	6.93
		minDCF	0.40	0.42	0.36	0.32
	TI	EER	11.63	15.29	11.06	9.22
		minDCF	0.49	0.60	0.46	0.42

VI. Conclusions and Future Works

We focused on the short-duration problems on speaker verification, and defined a new concept – the sub-3sec problem. Then 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.

Future work includes further dividing the subsets within the Sub3Vox, and using this pipeline to organize more and larger TD-SV corpora, such as from VoxCeleb2 and VoxBlink. This will contribute to the development of research on short-time text-dependent speaker verification.

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] S. Wang, Z. Chen, K. A. Lee, Y. Qian, and H. Li, "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning," *arXiv preprint arXiv:2407.15188*, 2024.

[3] M.-W. Mak and J.-T. Chien, *Machine learning for speaker recognition*. Cambridge University Press, 2020.

[4] 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.

[5] M. Hébert, "Text-dependent speaker recognition," *Springer handbook of speech processing*, pp. 743–762, 2008.

[6] 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.

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

[8] 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.

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

[10] 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.

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

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

[13] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, "Meta-learning without memorization," *arXiv preprint arXiv:1912.03820*, 2019.

[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 *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 *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 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 *Annual Conference of the International Speech Communication Association*, 2012.

[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 *Odyssey*, 2018, pp. 386–392.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Annual Conference of the International Speech Communication Association*, 06 2017.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Annual Conference of the International Speech Communication Association*, 06 2018.

[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 *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 10 271–10 275.

[22] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances." in *Annual Conference of the International Speech Communication Association*, 08 2011.

[23] 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.

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

[25] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[26] 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 *Annual Conference of the International Speech Communication Association*, 09 2015.

[27] 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.

[28] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[34] 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 *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[35] 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.