Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition

Anonymous submission to INTERSPEECH 2023

Abstract

The success of these systems heavily depends on large training datasets collected under real-world conditions. While common languages like English or Chinese have vast available datasets, low-resource ones like Vietnamese remain limited. Existing Vietnamese speaker datasets are typically limited in scale or collected under constrained conditions. This paper presents a large-scale spontaneous dataset gathered under noisy environments, with over $87,000$ utterances from $1,000$ Vietnamese speakers of many professions, covering 3 main Vietnamese dialects. To build the dataset, we propose a sophisticated construction pipeline that can also be applied to other languages, with efficient visual-aided processing techniques to boost data precision. With the state-of-the-art x-vector model, training with the proposed dataset shows an average EER absolute and relative improvement of 5.48% and 41.61% when compared to the model trained on VLSP 2021, a publicly available Vietnamese speaker dataset.

Index Terms: speaker recognition, speaker verification, vietnamese dataset

1. Introduction

Speaker recognition is the task of identifying or verifying the identity of an individual based on their speech segments. After decades of research, the development of speaker recognition systems have seen signification advancements, and have been deployed to various practical application, such as access control, forensic investigations, and personalized services. However, all such fields require speaker recognition systems to excel under real-world conditions, where there exist multiple uncertain factors, including environmental noises, channel and microphone effects; or intrinsic variations such as speaking styles, accent, or physiological status.

The development of speaker recognition systems highly depends on the size and reliability of the data. In today's world, there have been some available datasets specified for this task but most of them are in common-used languages such as English or Chinese. Some of the successful and freely available datasets include VoxCeleb [1], the VoxCeleb2 [2], the Speakers In the Wild (SITW) dataset [3] or the CN-Celeb [4] dataset. However, with low-resource languages, including Vietnamese, possessing a great number of speakers is still necessary. Therefore, having the same number of celebrities as CN-Celeb can be considered challenging due to the low population and number of celebrities compared to other big countries. Currently, the two publicly available Vietnamese datasets, called ZaloAI [5] and VLSP 2021 [6, 7], perform well for this speaker verification task, with an Equal Error Rate (EER) of 3.115%. However, the reliability of these datasets is still a question mark. The data-building pipeline of these datasets is insufficient since it lacks visual-aided pre-processing techniques applied in VoxCeleb or CN-Celeb. This can cause data mislabeling on both the train set and the test set.

As previous studies have not completely solved the problem of a low-resource language dataset with high reliability, in this study, we propose Vietnam-Celeb - a well-built dataset for Vietnamese speaker recognition. This dataset consists of more than 87,000 utterances from 1,000 Vietnamese celebrities. We take Youtube and Tiktok to be ideal places to take celebrities utterances. To preserve the precision of the data, we combine effective data-collecting strategies with sufficient visual-aided data processing techniques, such as face tracking, face verification, and active speaker verification. Furthermore, to our knowledge, this is the first Vietnamese speaker dataset that includes information about the speaker's gender and his/her dialect. These additional features help our data to be much more useful for the Vietnamese speaker verification model and any further speech processing system that involves Vietnamese.

The rest of this paper is organized as follows: Section 2 describes our data collecting and pre-processing pipeline, while Section 3 discusses the process and result of building the new dataset. Our experimental results are provided in Section 4. Finally, we draw conclusions in Section 5.

2. Dataset collection & construction

2.1. Overall pipeline

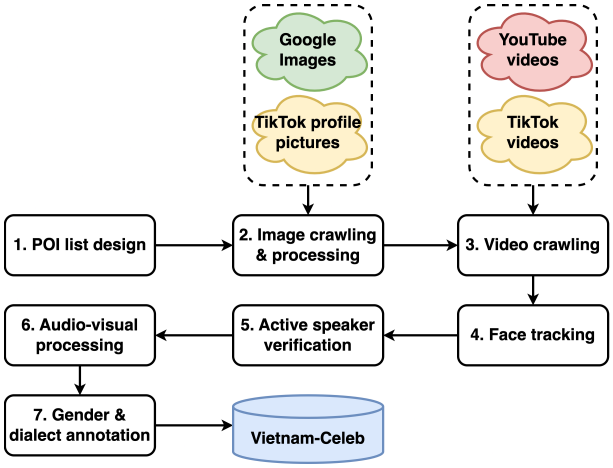


Figure 1: *Dataset collection & construction pipeline.*

We propose a speaker dataset-building pipeline describing

our multi-stage approach to collecting a large speaker recognition dataset. Figure 1 illustrates the stages in the pipeline. The proposed pipeline can also be applied to other languages, as the chosen data sources are common for many languages. Starting from a list of POIs, images and videos are crawled for the designated POI, then we apply processing procedures and visual-aided filtering, such as S^3FD [8] for face tracking and TalkNet [9] for active speaker verification to obtain a thousand speakers with different utterances.

As low-resource languages like Vietnamese data are more challenging to collect compared to Chinese (CN-Celeb) or Multilingual data (VoxCeleb), several modifications are added to our proposed pipeline. The following sections will discuss the main stages with significant enhancements compared to the pioneering works of VoxCeleb and CN-Celeb.

2.2. Designing list of Persons of Interest (POI)

We manually select a list of Vietnamese celebrities from Wikipedia. The list includes 700 celebrities, which covers a wide range of professions, including actors, singers, business people, and athletes. We also select manually 200 TikTok channels of Vietnamese celebrities. Additionally, we also choose various YouTube playlists to obtain additional speakers. Only playlists with video titles named in the same patterns are chosen so that we can use simple rules to easily mark out the name of POI. At last, we obtain a list of about 1300 POI in total. However, designing the POI list in this way will result in duplicated speakers. We will deal with this problem at a later stage.

2.3. Image crawling and processing

For image crawling, we use different strategies for different data sources. In the case of POI coming from TikTok, we download the TikTok profile picture for each POI. For the rest of the POI list, we download 100 images from Google Images for each POI. When searching for images, we add the phrase *portrait of* to the query to make sure the images contain the face of the target POI.

After crawling the images, we begin to process them. For the images crawled from Google, we have to filter out irrelevant images for each POI. For each downloaded image, we perform face detection using RetinaFace [10], then we use ArcFace [11] to extract the face embeddings for each detected area. With a list of face embeddings for a POI, we perform k-means clustering and for each cluster, we first remove the image samples that fall into the following categories:

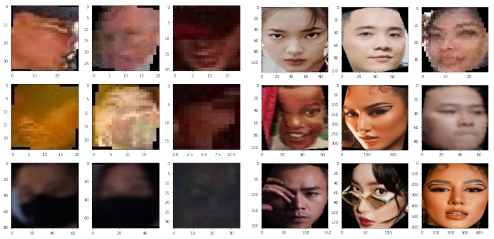
- Image size is too small (below 64x64).
- Average cosine similarity score between the face embeddings of that image and other embeddings in the same cluster is lower than a pre-defined threshold.

Finally, we remove the clusters having below 10 valid images. Figure 2a and Figure 2b illustrate two invalid clusters that will be removed. After performing the processing steps, the average number of images for each POI is 33.

2.4. Video crawling

The two sources where we collected our data are YouTube and TikTok. We choose to crawl 100 top videos for each TikTok POI. For the POI taken from YouTube video titles, we download all videos of the playlists where the POI names appear in the titles.

For the list of celebrities obtained from Wikipedia, we apply YouTube Advanced Search Operators when searching for videos to ensure the POI's name and the keyword *interview* appear in the video title, which should increase the chance that the person is speaking in the video. With each search result of a POI, we collect only the top 20 videos, as we find that in many cases, the videos after the top 20 are mostly irrelevant, which either do not contain the desired keyword or POI's name.



(a) *Invalid cluster containing small images.* (b) *Invalid cluster containing false images.*

Figure 2: *Invalid cluster containing false images.*

2.5. Audio-visual processing

After obtaining cleaned video segments of every POI from stage 5, several refinement steps are conducted to ensure the validity of the collected utterances.

The first step is removing duplicated utterances, as the collected videos may contain re-uploads. We detect duplicated utterances by using a speech representation model pre-trained on Vietnamese data - Wav2Vec2 [13]. We then compute the cosine distance between every pair of utterances in the dataset and keep only one of them if their distance is below a very conservative threshold.

The second step is to remove invalid speakers. A speaker is considered invalid if the cross-similarity score among his utterances is low. The cross-similarity score is the average cosine score of the pairwise similarity matrix between the utterances.

The third step is removing noisy utterances using outlier detection. Let Q1 and Q3 be the first quartile and third quartile of the set of average cosine similarity score $a_i = \frac{1}{n}\sum_{j=0,j\neq i}^{n-1} \boldsymbol{S}_{i,j}$. By using interquartile range (IQR), the valid utterance range $[a_{min}, a_{max}]$ of set $\boldsymbol{a}$ can be determined in (1) and (2). Utterances having similarity scores a_i outside of the valid range $[a_{min}, a_{max}]$ are considered to be noisy utterances and will be removed from the dataset.

$$a_{min} = Q1 - 1.5 * IQR;\ a_{max} = Q3 + 1.5 * IQR \quad (1)$$

$$IQR = Q3 - Q1 \quad (2)$$

The last step is merging duplicated speakers, as there are likely duplicated speakers in the POI list. The merging is based on the average audio similarity scores of utterances and the average visual similarity scores between speakers.

2.6. Annotating speaker genders and Vietnamese dialects

The last step in our pipeline is creating the gender and dialect labels for each speaker. Dialect is an important factor when designing a Vietnamese speech dataset, as speeches uttered in different dialects have very different characteristics. There are

Table 1: *Comparison of some existing speaker recognition datasets*

Name	Environment	Language	Data Sources	Visual-aided	# of Spks	# of Utters	# of Hours
SITW [3]	interview	English	open-source media	No	299	2,800	-
VoxCeleb1 [1]	mostly interview	Mostly English	YouTube	Yes	1,251	153,516	352
VoxCeleb2 [2]	mostly interview	Multilingual	YouTube	Yes	6,112	1,128,246	2,794
CN-Celeb1 [4]	multi-genre	Chinese	Bilibili	Yes	1,000	130,109	274
CN-Celeb2 [12]	multi-genre	Chinese	multi-media	Yes	2,000	529,485	1,090
VLSP 2021 [6]	multi-genre	Vietnamese	ASR datasets	No	1,305	31,600	41
Vietnam-Celeb	**multi-genre**	**Vietnamese**	**YouTube, TikTok**	**Yes**	**1,000**	**87,140**	**187**

three main dialects in the Vietnamese language: Northern dialect, Central dialect, and Southern dialect. Hence, theoretically, we can probably obtain the dialect/accent labels by processing information about the birthplaces of celebrities. However, there is one problem with Vietnamese celebrities, which is that they tend to switch to a more common dialect (Northern or Southern dialect) when speaking in public for easier understanding with the audience.

To address this problem, we decide to manually create gender and dialect labels for every speaker. We sample 5 utterances for each speaker, annotated by 3 annotators through voting. The three annotators whom we choose come from 3 different Vietnamese regions, Northern, Central, and Southern regions so they have a high chance of correctly classifying speaker dialects.

3. The Vietnam-Celeb dataset

After going through the pipeline, we obtain the Vietnam-Celeb dataset, which consists of 1,000 speakers and more than 87,000 utterances. The total duration of the dataset is 187 hours. Our data covers a wide range of challenging scenarios, including interviews, podcasts, game shows, talk shows, and other types of entertainment videos. The audio samples also represent real-world conditions, where there are various types of noises, such as background chatting, music, and cheers. The sections below will discuss several statistics of Vietnam-Celeb. Vietnam-Celeb is publicly available for researchers, published under an anonymous GitHub repository [1].

3.1. Utterance length distribution

Table 2: *Utterance length distribution*

Length (s)	# of Utterances	Proportion (%)
<2	4,889	5.8%
2-5	**38,159**	**43.8%**
5-10	23,112	26.4%
10-20	14,155	16.1%
20-30	4,518	5.2%
>30	2,297	2.7%

Table 2 shows the utterance length distribution of the dataset. Short utterances make up a high amount of our data, which represents the audio the real-world speaker recognition task, where the audio inputs are mostly short.

3.2. Gender and dialect

The proposed dataset is gender-balanced, with 552 male speakers, accounting for 55.2% of the speakers. For the dialect distribution, Table 3 shows the statistics of dialect in each data source. As mentioned in the previous section, Vietnamese celebrities tend to switch to their northern or southern dialect when speaking, which explains the small number of central-dialect speakers in our dataset.

Table 3: *Dialect statistics of Vietnam-Celeb*

Dialect	Northern	Central	Southern
YouTube			
# of Spks	409	32	383
# of Utters	31,220	2,141	27,284
# of Hours	61.13	5.15	62.81
TikTok			
# of Spks	102	8	64
# of Utters	15,902	1,009	9,584
# of Hours	33.64	2.31	22.33

3.3. Comparison with existing datasets

Table 1 shows the comparison of several speaker recognition datasets with Vietnam-Celeb. Compared to the initial builds of VoxCeleb and CN-Celeb, Vietnam-Celeb has a roughly smaller number of speakers, as this is the first Vietnamese speaker dataset that employs visual-aided processing in the building pipeline. Compared to another publicly available Vietnamese speaker dataset - VLSP 2021 - Vietnam-Celeb is a more complex and refined dataset:

- The number of utterances and the total duration of Vietnam-Celeb is significantly higher than those of VLSP 2021.
- Vietnam-Celeb covers a wider variety of speech scenarios and can represent real-world noisy speech conditions.
- Vietnam-Celeb employs visual-aided processing to ensure the correctness of the data. VLSP 2021 is built using only a speaker recognition system in the pipeline.
- Vietnam-Celeb includes gender and dialect labels for all speakers, which is crucial in building a Vietnamese speech dataset.

3.4. Final dataset

Table 4: *Statistics of Vietnam-Celeb subsets*

Subset	# of Spks	# of Utters	# of Utter Pairs
Vietnam-Celeb-T	880	82,907	-
Vietnam-Celeb-E	120	4,207	55,015
Vietnam-Celeb-H	120	4,217	55,015

For later research to experiment with our dataset, we have split the training set and two test sets from Vietnam-Celeb. Ta-

[1] https://github.com/Vietnam-Celeb/Vietnam-Celeb

ble 4 shows the statistics of each set. To obtain the test sets, we sampled 120 speakers from the data, with consideration to making sure the test data is gender-balanced and dialect-balanced. Additionally, inspired by the work of VoxCeleb1 [1] when creating the test sets, 120 speakers are chosen among the speakers who have the highest speech similarity scores and visual similarity scores. As the number of central-dialect speakers is very low, we took all of them into building the test sets. With the obtained speakers, we then created two test sets:

- **Vietnam-Celeb-E**: An easy test set of Vietnam-Celeb. The negative pairs in the set - pairs of different speakers - are sampled randomly.
- **Vietnam-Celeb-H**: A hard test set of Vietnam-Celeb, which takes gender and dialect information into account. For creating the negative pairs, we make sure that two speakers in each pair have the same gender and dialect labels.

4. Experiments

4.1. Model architecture

We chose the ECAPA-TDNN architecture [14] for our experiments. The input to the network is 80-dimensional MFCCs from a 25 ms window with a 10 ms frameshift. To convert the frame-level features into utterance-level features, attentive statistics pooling (ASP) is used. The utterance-level features are then passed through a fully connected layer (FC) to produce speaker embeddings. Finally, the model is optimized using AAM-Softmax Loss [15].

4.2. Experimental setup

The evaluation metric to be used in experiments is Equal Error Rate (EER). To assess the performance of our models trained on Vietnam-Celeb-T, we compare them with the EPACA-TDNN models trained on VoxCeleb and VLSP 2021, which follow the same model configurations.

For the VoxCeleb ECAPA-TDNN model, we used the pre-trained model from [16]. The models trained with Vietnam-Celeb-T and VLSP 2021 are trained from scratch using the Adam optimizer for 100 epochs, with a batch size of 100 and an initial learning rate of 5e-4, decayed linearly.

We also experimented with using VoxCeleb as the pre-training dataset and fine-tuning the model with either VLSP 2021 or Vietnam-Celeb. For fine-tuning, we used the same training configurations as above except that we set the number of epochs to 30. The trained models are denoted as follows:

- **Vox**: The VoxCeleb pre-trained model.
- **VLSP**: The model trained on the VLSP 2021 training set.
- **Vietnam-Celeb**: The model trained on Vietnam-Celeb-T.
- **Vox + VLSP**: The VoxCeleb pre-trained model which is fine-tuned with the VLSP 2021 training set.
- **Vietnam-Celeb**: The VoxCeleb pre-trained model which is fine-tuned with Vietnam-Celeb-T.

4.3. Experimental results

We evaluate the models on 3 test sets: Vietnam-Celeb-E, Vietnam-Celeb-H, and VLSP 2021 to test the proposed dataset capabilities when being used as either a standalone training set or as a supporting dataset served for fine-tuning.

Table 5: *EER (%) results on Vietnam-Celeb test sets (lower is better)*

Model	Vietnam-Celeb-E	Vietnam-Celeb-H
Vox	13.19	16.52
VLSP	11.58	14.3
Vietnam-Celeb	**6.31**	**8.62**
Vox + VLSP	10.92	13.49
Vox + Vietnam-Celeb	7.33	9.37

We first compare the results of models on the two evaluation sets of Vietnam-Celeb. Table 5 shows the experimental results. In every case, the EER on Vietnam-Celeb-E is lower than that on Vietnam-Celeb-H, as the latter contains hard negative pairs. The *Vox* model performs the worst on the two sets. *Vietnam-Celeb* also outperforms *VLSP* with an average relative EER improvement of 41.61%. In cases of the fine-tuned models, while *Vox + VLSP* achieves better results compared to *VLSP*, using VoxCeleb as the pre-trained model does not improve the results when fine-tuning with Vietnam-Celeb-T. These results emphasize the capability of Vietnam-Celeb when used as a standalone training set.

Table 6: *EER (%) results on the VLSP test set (lower is better)*

Model	VLSP 2021 test set
Vox	5.92
Vox + VLSP	4.57
Vox + Vietnam-Celeb	**3.69**

To assess the supporting capability of Vietnam-Celeb, we experimented with different models on the VLSP 2021 test set, as shown in Table 6. The overall results are better than those of Vietnam-Celeb evaluation sets, which shows that this test set is much less challenging compared to Vietnam-Celeb test sets. While *Vox* has already achieved a decent EER, of about 5.92%, *Vox + Vietnam-Celeb* illustrates the lowest EER, of about 3.69%, which demonstrates the highly efficient supporting ability of our proposed dataset when used for fine-tuning.

5. Conclusion

This paper presents Vietnam-Celeb - a new large-scale dataset for Vietnamese speaker recognition. The dataset is built with a sophisticated data construction pipeline, with efficient visual-aided processing techniques to preserve the precision of the data. Vietnam-Celeb contains over 87,000 utterances from 1,000 Vietnamese celebrities and is the first Vietnamese speaker recognition dataset to come with gender and dialect labels. To conduct our experiments, we apply different training strategies with 3 datasets, VoxCeleb, VLSP 2021, and Vietnam-Celeb. The results show that in every case, models trained with our proposed dataset outperform those trained with a publicly available Vietnamese dataset, VLSP 2021 where they show an average EER absolute and relative improvement of 5.48% and 41.61%, respectively. We believe that this dataset will be highly useful for the research community on Vietnamese. In future works, we aim to extend Vietnam-Celeb with coverage to more various speech environments, as well as improve the current data-building pipeline, which can be to add a manual annotation step to enhance the data quality.

6. References

[1] 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.

[2] J. Son Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv e-prints*, pp. arXiv–1806, 2018.

[3] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[4] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[5] "Zalo-AI challenge," https://challenge.zalo.ai/.

[6] V. T. Dat, P. V. Thanh, and N. T. T. Trang, "VLSP 2021 - SV challenge: Vietnamese speaker verification in noisy environments," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 1, 2022.

[7] D. V. Thanh, T. P. Viet, and T. N. T. Thu, "Deep speaker verification model for low-resource languages and Vietnamese dataset," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China: Association for Computational Lingustics, 11 2021, pp. 442–451. [Online]. Available: https://aclanthology.org/2021.paclic-1.47

[8] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single shot scale-invariant face detector," *CoRR*, vol. abs/1708.05237, 2017. [Online]. Available: http://arxiv.org/abs/1708.05237

[9] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[10] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," *CoRR*, vol. abs/1905.00641, 2019. [Online]. Available: http://arxiv.org/abs/1905.00641

[11] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *CoRR*, vol. abs/1801.07698, 2018. [Online]. Available: http://arxiv.org/abs/1801.07698

[12] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: multi-genre speaker recognition," *CoRR*, vol. abs/2012.12468, 2020. [Online]. Available: https://arxiv.org/abs/2012.12468

[13] 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.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[15] J. Son Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv e-prints*, pp. arXiv–2003, 2020.

[16] R. K. Das, R. Tao, and H. Li, "HLT-NUS SUBMISSION FOR 2020 NIST CONVERSATIONAL TELEPHONE SPEECH SRE," *arXiv preprint arXiv:2111.06671*, 2021.