



# A Novel Phoneme-based Modeling for Text-independent Speaker Identification

Xin Wang<sup>1</sup>, Chuan Xie<sup>1</sup>, Qiang Wu<sup>1</sup>, Huayi Zhan<sup>1</sup>, Ying Wu<sup>2</sup>

<sup>1</sup>Changhong AI Lab (CHAIR), Sichuan Changhong Electronics Holding Group Co., Ltd.

<sup>2</sup>Electrical Engineering and Computer Science, Northwestern University, US

{xin13.wang, chuan.xie, qiang4.wu, huayi.zhan}@changhong.com, yingwu@ece.northwestern.edu

## Abstract

Text-independent speaker identification attracted growing attention while it remains challenging to extract speaker-specific features from a speech with arbitrary content. End-to-end systems trained with utterance-level features suffer from performance degradation caused by speech content variation. To address this issue, this paper proposes a novel phoneme-based approach with the following key features: first, it restricts the variety of speech content by splitting each utterance into a set of phoneme segments and develops the phoneme-constrained models to extract segment-level embeddings of speakers; second, it leverages a soft-voting mechanism with mono-phonemic thresholds and weights to combine the results of different phonemes. Experimental results on AISHELL and ASRU2019 datasets show that the proposed approach is effective and robust, which outperforms the state-of-the-art methods in both EER and accuracy, especially with a larger phonemic mismatch between the enrollment and test utterances. In addition, the proposed system is efficient that can be trained well on a small-scale dataset.

**Index Terms:** phoneme-based models, text-independent speaker identification, segment-level feature extraction

## 1. Introduction

Speaker identification, as one of the most crucial methods to distinguish people, has attracted a wide range of attention. Compared with the text-dependent speaker identification systems [1, 2] which require a fixed or constrained text phrase, the text-independent approaches [3] can work on arbitrary speech content. In the last decade, with the success of Gaussian mixture models (GMMs) [3, 4] and deep neural networks (DNNs) [5, 6, 7, 8, 9] on extracting speaker representations, researchers have made significant progress in text-independent tasks. Nevertheless, how to capture speaker identity-related information from a speech signal with arbitrary text remains challenging. On the one hand, some studies have proved that the distribution of speaker information is not uniform on speech signals [10, 11, 12]. On the other hand, a sentence can contain any combination of words and phrases that create a huge text space. In order to reduce the impact of various contents in the speech signals, most of the end-to-end methods improve the performance by expanding the amount of training data to let the models learn as much text as possible. Although DNNs are powerful, speech content variation increases the complexity of network training, and it is still difficult to extract speaker identity information from a speech signal with text that has not been “seen” in the training set. Furthermore, the phonetic mismatch between enrollment and test utterances also leads to performance degradation [13].

Previous works have tried to solve this issue in several ways. One is to introduce the multitask framework to encourage or suppress phonetic information [14, 15, 16, 17]. Specifically,

a multitask model that combines the speaker vectors from an x-vector network and the phonetic vectors from an automatic speech recognition (ASR) network was proposed by [14] in order to use the phonetic information to adapt the speaker embeddings. Similarly, a coupled network was designed to fuse the phonetic information from ASR and the acoustic features [18]. Authors in [15, 16] took advantage of adversarial approaches to suppress the impact of nuisance phonemic variation. The method in [17] exploited the phoneme-aware attentive pooling and used the phoneme posterior from the subnet to capture the phonetic information for text-dependent task. Another way to handle this issue is to revise the extraction procedure. For example, a mixture of Gaussians substitutes a single Gaussian in the total variability model based on the assumption that different phonemes have different priors [19]. The posteriors obtained by an ASR-DNN are used as the replacement of the Universal Background Model for i-vector extraction [20]. And a fusion of the PLDA score with the phonetic information distance between enrollment and test utterances is proposed and attached to the i-vector system in order to compensate the phonetic mismatch [13]. Besides, an adaptive network with input-dependent kernels divided the utterances along the time axis to reflect the phonemes-varying characteristic [21]. Although these approaches can mitigate the impact of content variation to some extent, the features of speech content and speaker identification are still coupled.

Inspired by ASR that usually leverage a lexicon to map the phonemic sequences into words, we divide the words, phrases and sentences into phoneme segments and classify them into the corresponding phonemic categories. Therefore, we can explicitly remove the impact of text variation by converting the text-independent task into a set of text-dependent sub-tasks. The researchers in [22, 23, 24, 25, 26, 27] used GMMs to investigate the effectiveness of phoneme-constrained or syllable-constrained models. However, these text-constrained models were built with frame-level features that result in marginal improvement in accuracy. Differently, our proposed system leverages segment-level features by aggregating frames across the phoneme duration and takes advantage of the temporal relation. Intuitively, the pronunciation of a specific phoneme is more like a stochastic process in cross-time rather than a random variable. The information carried by an individual frame is insufficient for neither speaker identification nor phoneme recognition. On the contrary, the segment-level features learned from complete phoneme duration can provide richer information that is more discriminative.

The major contributions of this paper are as follows: First, we propose a phoneme-based speaker identification system that can explicitly eliminate the impact of speech content variation by dividing utterances into a set of phoneme segments. Second, we develop an effective method to integrate phoneme-constrained models into a speaker identification system with

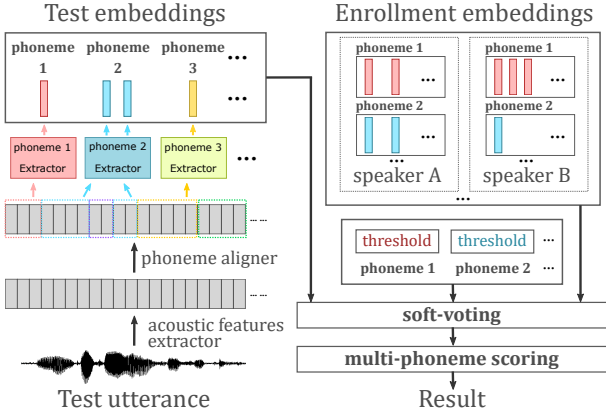


Figure 1: The overall architecture of our proposed system.

a superior performance by employing soft-voting mechanisms and a set of mono-phonemic thresholds; Finally, experiments on AISHELL and ASRU2019 datasets demonstrate that the proposed approach is efficient to be trained on a small-scale dataset and is robust with a larger phonemic mismatch between enrollment and test utterances.

## 2. Proposed Phoneme-based Framework

### 2.1. Phoneme alignment and segmentation

To detect the phoneme categories and boundaries, a two-stage method is employed as follows. First, we obtain the transcription of utterances and convert graphemes to phonemes by using a predefined grapheme-to-phoneme dictionary. Second, we use a forced alignment module to find the proper boundary of each detected phoneme. Then the acoustic features of utterances are split into phonemic segments based on the boundaries provided by the aligner.

### 2.2. Phonemic speaker embeddings extraction

A convolutional neural network (CNN) is employed to extract segment-level speaker embeddings for each phoneme class, which comprises three 1-dimensional CNN layers (kernel sizes 3 with 32, 64, 128 channels) and each of them is followed by a ReLU nonlinearity and a batch normalization layer. Then, an average pooling layer is used to aggregate frame-level features into a segment-wise embedding, which is connected to a fully connected layer. We feed the phonemic segments belonging to the same phoneme class into the corresponding phoneme-constrained model, and use the speaker labels as the ground-truth in training. Once the models are well-trained, the fully connected layer is removed and the outputs of the last layer are the phonemic speaker embeddings.

### 2.3. Mono-phonemic distance thresholds

A test utterance is split into  $N$  phonemic segments and represented as a set of phonemic embeddings  $\{e_1, \dots, e_N\}$  with phoneme class  $\{p_1, \dots, p_N\}$ . The  $n^{\text{th}}$  phonemic segment embedding from the test utterance assigned to phoneme class  $p_n$  is denoted as  $e_n(p_n)$ . First of all, we define the distance between any two embeddings as  $D(e_i, e_j) = (1 - \cos \theta)/2$ , where  $\theta$  is the angle between embedding  $e_i$  and  $e_j$ . Note that we only compare the phonemic embeddings in the same phoneme classes. That is, for the test embedding  $e_n(p_n)$ , only the enroll-

ment embeddings in phoneme class  $p_n$  are considered. We use  $\epsilon(P, Y, I)$  to denote an enrollment embedding that is assigned to the phoneme class  $P$  with index  $I$  and labeled as speaker  $Y$ . For simplicity, we denote the distance between the test embedding  $e_n(p_n)$  and the enrollment embedding  $\epsilon(p_n, Y, I)$  as  $d_n(p_n, Y, I)$ . We design a distance threshold for each phoneme class to reject embeddings that is too far from the test sample in the feature space and collect the near embeddings to build the valid embeddings set  $\Omega$ . For example, if the distance  $d_n(p_n, Y, I)$  is smaller than the corresponding threshold  $t(p_n)$ , the enrollment embedding  $\epsilon(p_n, Y, I)$  should be collected in the valid set  $\Omega_n(p_n)$  and excluded otherwise. It can be described by:

$$\Omega_n(p_n) = \{\epsilon(p_n, Y, I) | d_n(p_n, Y, I) < t(p_n)\}. \quad (1)$$

The threshold of a certain phoneme class is fixed after training, while the valid set is computed dynamically for each test embedding. In order to reduce the calculating cost, we can limit the maximum number of embeddings in a valid set to be  $k$  that is similar to the k-nearest neighbors algorithm [28].

### 2.4. Soft-voting mechanism

Once the valid set for a test embedding is acquired, we employ the soft-voting mechanism to make predictions. The vote from a test segment  $e_n(p_n)$  to each enrollment in the valid set  $\Omega_n(p_n)$  is computed by softmax as:

$$v_n(p_n, Y, I) = \frac{\exp(-\frac{d_n(p_n, Y, I)}{\tau})}{\sum_{\epsilon(p_n, \tilde{Y}, \tilde{I}) \in \Omega_n(p_n)} \exp(-\frac{d_n(p_n, \tilde{Y}, \tilde{I})}{\tau})}, \quad (2)$$

where  $v_n(p_n, Y, I)$  represents the vote obtained by the enrollment embedding  $\epsilon(p_n, Y, I)$ , and  $\tau$  is a non-negative temperature. Each test embedding  $e_n$  has one vote that can be apportioned to every enrollment in the corresponding valid set  $\Omega_n$  as:

$$v_n(p_n) = \sum_{\epsilon(p_n, Y, I) \in \Omega_n(p_n)} v_n(p_n, Y, I) = 1. \quad (3)$$

If an enrolled speaker has over one embedding in the valid set, each of them can acquire votes independently. For a given speaker  $y$ , the total votes obtained from a test embedding  $e_n$  is:

$$v_n(p_n, Y = y) = \sum_I v_n(p_n, Y = y, I). \quad (4)$$

In our design, if a phoneme category is repeated in a test utterance, each occurrence of this phoneme is regarded as an individual phoneme so that it can vote independently.

### 2.5. Multi-phoneme scoring

After obtaining votes from individual phonemic segments in the test utterances, the next step is to combine the votes and calculate a score for each enrolled speaker. Intuitively, as the discriminative capabilities of different phoneme classes are unequal, different priors should be assigned to every phoneme class as a weight coefficient in the combination. For a test utterance consisting of  $N$  phonemes, the score obtained by an enrolled speaker  $Y$  is the weighted sum of votes from  $N$  phonemes and the result can be normalized as:

$$s(Y) = \frac{\sum_{n=1}^N v_n(p_n, Y) w(p_n)}{\sum_{n=1}^N w(p_n)}, \quad (5)$$

where  $w(p_n)$  is the weight coefficient assigned to the phoneme class  $p_n$ . Thus, it is feasible to calculate an optimal threshold to reject non-target speakers in utterance level.

Table 1: Performance comparison on the AISHELL-1 dataset in different tasks.

Model	EER%		Accuracy%	
	Random	Mismatch	Random	Mismatch
Thin ResNet-34 [7]	1.85	1.85	97.44	97.33
TDNN(x-vector) [5]	1.16	1.23	97.96	97.90
SEG+MT(segment-phoneme) [15]	1.39	1.31	97.04	97.12
SEG+ADV(segment-phoneme) [15]	1.39	1.23	97.25	97.46
SEG-MT(frame-phoneme) [16]	1.00	1.00	97.64	97.91
SEG-AT(frame-phoneme) [16]	1.58	1.43	95.42	95.85
Ours	<b>0.85</b>	<b>0.77</b>	<b>98.04</b>	<b>98.25</b>

Table 2: Performance comparison on the ASRU2019 dataset in different tasks.

Model	EER%		Accuracy%	
	Random	Mismatch	Random	Mismatch
Thin ResNet-34 [7]	1.79	1.81	82.42	82.09
TDNN(x-vector) [5]	2.10	2.12	78.92	78.51
SEG+MT(segment-phoneme) [15]	1.82	1.96	80.66	80.01
SEG+ADV(segment-phoneme) [15]	1.79	1.97	79.58	79.22
SEG-MT(frame-phoneme) [16]	1.73	1.93	82.91	82.10
SEG-AT(frame-phoneme) [16]	2.17	2.29	80.58	79.68
Ours	<b>1.56</b>	<b>1.51</b>	<b>84.59</b>	<b>85.25</b>

### 3. Experiments and Results

#### 3.1. Dataset

In this work, the experimental results are conducted using the Mandarin Speech Data from the ASRU 2019 Challenge dataset [29] and the AISHELL-1 dataset [30] with 16kHz sampled recordings.

Tasks on the ASRU2019 dataset are challenging because the short speech duration, approximately 2 seconds on average including silence, will decrease the accuracy of phoneme segmentation and increase the difficulty on phonemic feature extraction. The dataset consists of two subsets (“category 1” and “category 2”) and provides annotations of transcriptions. We use the “category 1” set as the training set, which contains 456,600 speech data from 1581 speakers. At the test stage, a subset of “category 2” that contains 100 speakers is used to evaluate the performance of speaker identification. The speakers in the test set are further split into the “target” and “non-target” groups with a 1:1 ratio, i.e., 50 enrolled speakers and 50 strangers.

AISHELL-1 is a widely-used multi-channel Mandarin speech dataset, including 141,600 utterances from 400 speakers. Following the official split, we use the speech utterances from 380 speakers in the training and development sets for model training and evaluate on the test set with 10 enrolled speakers (target) and 10 strangers (non-target).

On both datasets, we evaluate on two tasks: (1) in “Random” task, we enroll 100 utterances per target speaker which are selected randomly and the unseen utterances from both target and non-target speakers are used for evaluation; (2) in “Mismatch” task, we select 100 utterances with the largest number of phonemes for each target speaker to enroll and test on the unseen utterances from both target and non-target speakers.

#### 3.2. Experimental settings

We first utilize a dictionary to map the transcriptions provided by the datasets into pronunciation graphemes, followed by the

Table 3: The performance of the proposed phoneme-based approach and the baseline systems trained on a small dataset.

Model	EER%		Accuracy%	
	Random	Mismatch	Random	Mismatch
Thin ResNet-34 [7]	2.79	2.87	77.11	76.72
TDNN(x-vector) [5]	3.76	4.08	74.59	72.71
SEG+MT(segment-phoneme) [15]	3.05	3.39	76.12	75.14
SEG+ADV(segment-phoneme) [15]	3.39	3.65	77.15	76.10
SEG-MT(frame-phoneme) [16]	3.22	3.65	77.01	75.24
SEG-AT(frame-phoneme) [16]	4.94	5.44	68.68	67.40
Ours	<b>2.24</b>	<b>2.06</b>	<b>80.94</b>	<b>82.79</b>

Montreal Forced Aligner(MFA) [31] to obtain phoneme boundaries in utterances. Then we use a 25ms hamming window with step 10ms to generate frames from speech signals. Following [7], we convert the utterances into 257-dimensional spectrograms by employing a 512 point fast Fourier transform and a short-time Fourier transform. Each spectrogram is segmented into a set of phonemic fragments based on the phoneme boundaries provided by the aligner. Totally 63 phonemes are used, excluding silence. The models are implemented using Pytorch [32], and optimized by Adam optimizer [33] with an learning rate 0.001 and batch-size 256. The output embedding is 128-dimensional, and cross-entropy loss is employed. In our implementation, the hyper-parameter  $\tau$  is 1 and  $k$  is 10.

The baseline systems are the state-of-the-art methods, i.e., x-vector [5], Thin ResNet with GhostVLAD (Thin ResNet-34) [7] and four phoneme-invariant systems built on multitask framework composed of phoneme and speaker classification loss. The SEG-MT and SEG-AT leverage frame-level phonemic information [16], while the SEG+MT and SEG+ADV use segment-level phonetic content [15]. The baselines are also trained with 257-dimensional spectrograms. In the proposed method, each enrolled speaker is represented as a set of phonemic embeddings. For baseline systems, each enrolled speaker is represented as a speaker embedding. For AISHELL dataset, we generate 10 trials corresponding to the 10 enrolled speakers for each test utterance. Therefore, totally 61,760 trials (25,950 target and 35,810 non-target trials) are used to calculate the equal error rate (EER) and utterance-level threshold. Similarly, 50 trials for each test utterance in ASRU dataset are generated, and totally 2,016,450 trials (18,369 target and 1,998,081 non-target trials) are used. Then we predict the speaker identity of each test utterance and reject non-target speakers by using the utterance-level threshold, and calculate the identification accuracy.

#### 3.3. Evaluation Performance

To validate the effectiveness of the proposed approach, we first train the models on AISHELL training and development sets and Table 1 reports the EER and accuracy on test set. In general, the proposed approach outperforms all the baselines in both tasks, improving significantly compared with the methods without considering phoneme information, i.e., Thin ResNet-34 and TDNN, in terms of EER (0.85% vs 1.85%, 1.16%). As expected, in the “Mismatch” task, the EER of the TDNN system increases from 1.16% to 1.23% because of larger phonemic mismatch between enrollment and test, while the results of the phoneme-invariant systems become better. The EER of the proposed approach decreases from 0.85% to 0.77% and the performance gain benefits from the increasing number of phonemic embeddings in enrollment.

Then we use all utterances in the ASRU training set to build

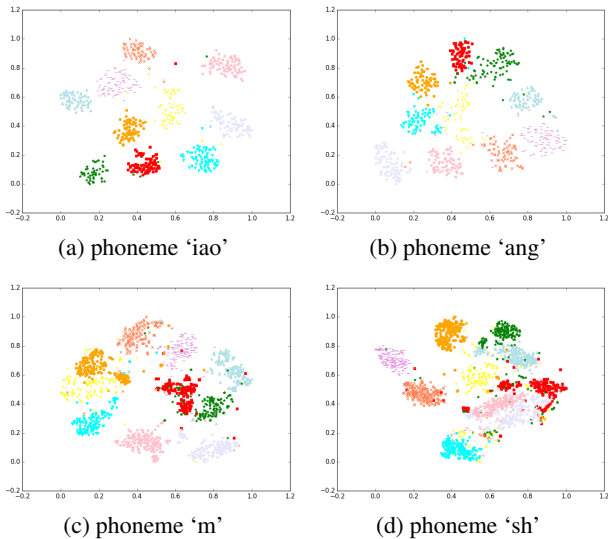


Figure 2: Visualization of phonemic embeddings in test set. Speakers are represented in different colors.

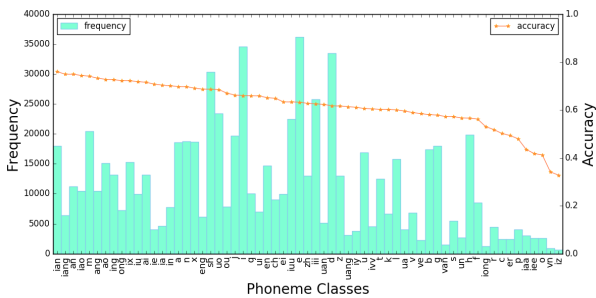


Figure 3: Comparison of occurrence frequency and identification accuracy of individual phoneme on the test set.

the systems and the results are summarized in Table 2. In both tasks, the proposed approach outperforms the baseline methods in both EER and accuracy. Specifically, the proposed approach achieves a better EER (1.56% vs 1.73%) and accuracy (84.59% vs 82.91%) than SEG-MT in the “Random” task, while the margins become larger (EER of 1.51% vs 1.93% and accuracy of 85.25% vs 82.10%) in the “Mismatch” task. Besides, all of the baselines perform worse in the “Mismatch” task, although the number of enrolled utterances is consistent with the “Random” task. It proves that the phonemic mismatch between enrollment and test utterances can deteriorate the performance of these systems. Unexpectedly, the performance degradation of the multitask systems which consider phonetic information in the training stage is more significant. On the contrary, our system performs even better in the “Mismatch” task, which means the proposed approach achieves robustness in the case of larger content differences between enrollment and test utterances.

Moreover, to evaluate the training efficiency of the systems, we use a small subset of the training data which contains 85,235 utterances from 200 speakers to feed the systems. Table 3 shows that the proposed approach trained on a small-scale dataset outperforms the state-of-the-art methods by a significant margin, especially in the “Mismatch” task. It indicates that our approach is more efficient in learning discriminative features.

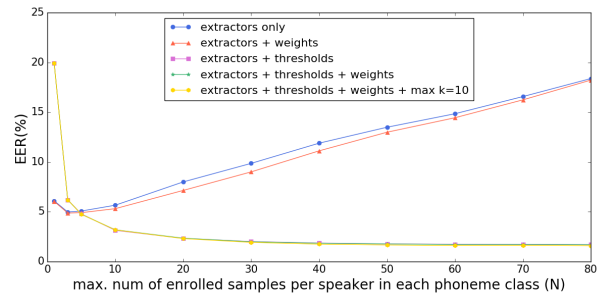


Figure 4: Ablation analysis for the proposed system.

### 3.4. Additional Experiments

To better understand the capability of the proposed approach, we conduct additional experiments on the ASRU2019 dataset. To illustrate the discrimination of phonemic embeddings, we visualize the phonemic embeddings by using t-distributed Stochastic Neighbor Embedding (t-SNE) [35]. In test set, 10 speakers are selected and 300 utterances are sampled randomly for each speaker. Figure 2 (a) and (b) show the distribution of embeddings extracted from two typical vowels/diphthongs, while (c) and (d) describe that of a nasal and a fricative, respectively. We also analyze the frequency of occurrence and the accuracy of speaker identification for each phoneme category in Figure 3. Unsurprisingly, most of the voiced phonemes perform better than the unvoiced, and it does not show a direct correlation between the frequency and accuracy. In a further investigation, we study the effects of each component in the proposed system and explore the relations between the number of enrolled phonemic segments and the system performance. As the quantities of enrolled samples in different phoneme categories are unbalanced, we use a maximum number  $N$  to limit the amount of enrolled samples per speaker in every phoneme class. The test data is consistent with the “Mismatch” task, and we randomly select  $N$  (or less than  $N$  if insufficient) samples per target speaker for enrollment in each phoneme class. Figure 4 shows that the proposed system benefits a lot from the mono-phonemic thresholds and there is a strong correlation between identification performance and the number of enrolled samples. With thresholds, more enrolled samples can contribute to better performance in general. Nevertheless, the trend becomes flat when  $N$  is larger than 10. After removing mono-phonemic thresholds, as the number of enrolled samples increases, the performance turns worse. The major reason might be that some enrolled speakers can obtain votes though their samples are far from the test embeddings in the feature space.

## 4. Conclusions

In this paper, we have proposed a phoneme-based approach to reduce the impact of speech content variation on speaker identification tasks by explicitly converting the text-independent utterance into a set of text-constrained segments, i.e., phonemes, and extracting segment-level speaker embeddings across the phoneme duration. Experimental results on AISHELL and ASRU2019 datasets show that the proposed approach achieves state-of-the-art performance of speaker identification, and it has better robustness and efficiency than previous methods. In future work, we plan to investigate multilingual phonemes and explore a more effective method of phoneme detection.

## 5. References

- [1] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP*, 2014, pp. 4052–4056.
- [2] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *INTERSPEECH*, 2015, pp. 185–189.
- [3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Speech Audio Process.*, 2011.
- [4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 4, pp. 1435–1447, 2007.
- [5] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, 2017, pp. 999–1003.
- [6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.
- [7] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP*, 2019, pp. 5791–5795.
- [8] M. India, P. Safari, and J. Hernando, "Double multi-head attention for speaker verification," in *ICASSP*, 2021, pp. 6144–6148.
- [9] Y. Wu, J. Zhao, C. Guo, and J. Xu, "Improving deep CNN architectures with variable-length training samples for text-independent speaker verification," in *Interspeech*, 2021, pp. 81–85.
- [10] M. Ajili, J. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Phonetic content impact on forensic voice comparison," in *2016 IEEE SLT*, 2016, pp. 210–217.
- [11] I. Magrin-Chagnolleau, J. Bonastre, and F. Bimbot, "Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods," in *EUROSPEECH*, 1995.
- [12] M. Ajili, J. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Phonological content impact on wrongful convictions in forensic voice comparison context," in *ICASSP*, 2017, pp. 2147–2151.
- [13] I. Viñals, A. Ortega, A. Miguel, and E. Lleida, "Phonetic variability influence on short utterances in speaker verification," in *IberSPEECH*, 2018, pp. 6–9.
- [14] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," in *INTERSPEECH*, 2018, pp. 2247–2251.
- [15] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernocký, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Interspeech*, 2019, pp. 1148–1152.
- [16] N. Tawara, A. Ogawa, T. Iwata, M. Delcroix, and T. Ogawa, "Frame-level phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances," in *ICASSP*, 2020.
- [17] Y. Liu, Z. Li, L. Li, and Q. Hong, "Phoneme-aware and channel-wise attentive learning for text dependent speaker verification," in *Interspeech*, 2021, pp. 101–105.
- [18] S. Zheng, Y. Lei, and H. Suo, "Phonetically-aware coupled network for short duration text-independent speaker verification," in *Interspeech*, 2020.
- [19] J. Ma, V. Sethu, E. Ambikairajah, and K. Lee, "Speaker-phonetic vector estimation for short duration speaker verification," in *ICASSP*, 2018, pp. 5264–5268.
- [20] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *ICASSP*, 2014.
- [21] S. Kim and Y. Park, "Adaptive convolutional neural network for text-independent speaker recognition," in *Interspeech*, 2021, pp. 66–70.
- [22] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained gaussian mixture models," in *ICASSP*, 2002.
- [23] W. D. Andrews, M. A. Kohler, J. P. Campbell, and J. J. Godfrey, "Phonetic, idiolectal and acoustic speaker recognition," in *Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 55–63.
- [24] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *INTERSPEECH*, 2005, pp. 2429–2432.
- [25] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *ICASSP*, 2009, pp. 4525–4528.
- [26] C. Vair, D. Colibro, F. Castaldo, E. Dalmaso, and P. Laface, "Loquendo - politecnico di torino's 2006 NIST speaker recognition evaluation system," in *INTERSPEECH*, 2007, pp. 1238–1241.
- [27] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, F. Grandas-Perez, and N. Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease," *Scientific Reports*, vol. 9, no. 1, p. 19066, 2019.
- [28] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine, 1951.
- [29] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.
- [30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*. IEEE, 2017, pp. 1–5.
- [31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldı," in *Interspeech*, 2017, pp. 498–502.
- [32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldı speech recognition toolkit," in *workshop on automatic speech recognition and understanding*, no. CONF, 2011.
- [35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.