2-2-5　言語識別・話者識別技術

2-2-5　Spoken Language and Speaker Recognition Technology

沈 鵬　Xugang Lu
Peng SHEN and Xugang LU

言語識別技術及び話者識別技術は、多言語音声翻訳システムの応用範囲を拡大する上で重要な技術である。本稿では、言語識別及び話者識別に関する我々の最新の研究成果を紹介する。言語識別については、短い発話に対する識別精度の改善手法及びクロスドメイン、クロスチャネルの問題に対してモデルの頑健性を改善する手法を紹介する。話者識別については、生成モデルと判別モデルの特徴を考慮したハイブリッドな学習手法により識別精度を改善する方法を紹介する。

Spoken language identification and speaker recognition are key technologies to enhance the application areas of multilingual speech translation systems. In this paper, we overview our latest studies of the two and the results we have achieved thus far. As for spoken language identification, we introduce techniques to improve the performance on short utterances and the model robustness for cross-domain/channel problems. As for speaker recognition, we introduce our proposed hybrid-learning method which takes into account the features of both generative and discriminative models.

1　Introduction

NICT is engaged in the development of highly practical, low-latency multilingual speech translation technologies that can be used in everyday life, such as public transport, business meeting, and international conferences. These technologies are essential for creating a society without language barriers in which the people of the world can communicate with each other without worrying about the differences in language or ability.

Unlike communication among humans, the current speech recognition-based systems require input direction- and language-switching each time before one speaks, due to the limitation of related technologies. These extra steps, which are not seen in communication between humans, hinder the application of speech-based systems. Developing speech context detection techniques, for example, language and speaker recognition, is essential for improving the usability of real-time multilingual speech translation systems.

As one of the most natural ways of communication, acoustic speech encodes various information. Besides linguistic information, non-linguistic or paralinguistic information is also important, for example, speaker information,

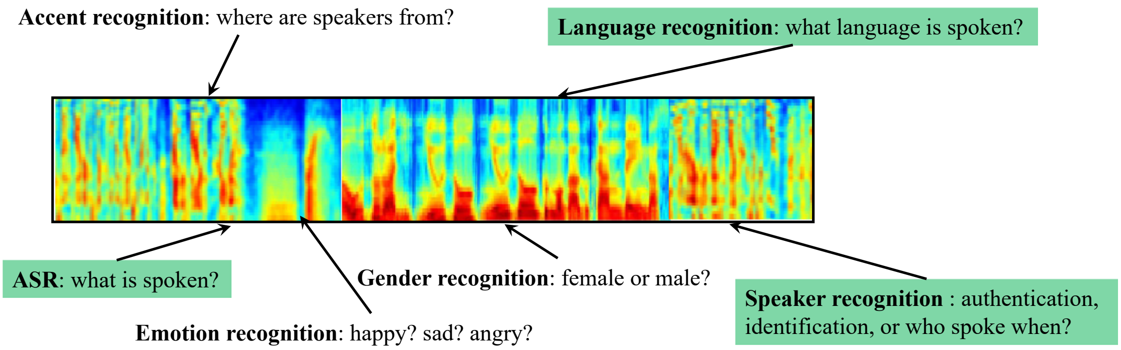

Accent recognition: where are speakers from?
Language recognition: what language is spoken?
ASR: what is spoken?
Gender recognition: female or male?
Speaker recognition : authentication, identification, or who spoke when?
Emotion recognition: happy? sad? angry?

Fig. 1　Information in acoustic signal and related speech technologies to decode them

linguistic information, emotion, accent, etc. In order to extract linguistic and paralinguistic information efficiently, several speech techniques have been developed. Fig. 1 illustrates the information encoded in acoustic speech and related speech technologies for decoding the underlying information.

As shown in this figure, in speech communication, besides the linguistic content transcribed by the automatic speech recognition (ASR) technique, we can see that several other non-linguistic patterns should also be identified by different techniques, for example, language recognition/identification, speaker recognition, emotion recognition, etc. Improving the detection and recognition accuracy of information is key to the success of the application of multilingual speech translation systems. In this paper, we explain our latest techniques for language and speaker recognition.

2 Spoken Language Identification

Spoken language identification (LID) is a task to determine which language is being spoken within a speech utterance [1]. Recently, LID techniques have been widely investigated and progressed. One conventional LID technique is the i-vector technique with conventional classifiers, such as support vector machine (SVM) and deep neural network (DNN). We also investigated the i-vector-based method to further improve the performance by using a local Fisher discriminative analysis and pair-wise distance metric learning [2][3]. Because the performance of i-vector techniques degrades on short utterance tasks, latest works use neural network-based techniques for building LID systems [4]-[6]. Although the neural network-based techniques showed their effectiveness on many LID tasks, in order to develop LID techniques for application, there are still some challenges that need to be overcome. In this paper, we introduce our works on two key challenges of the LID tasks: short utterance and cross-domain/channel problems.

2.1 Knowledge distillation for short utterance LID

LID technology is commonly used in the preprocessing stage of multilingual speech processing systems, such as spoken language translation and multilingual speech recognition. Traditional LID requires longer speech input to obtain better recognition performance, in real-time systems, the usage of longer speech causes delay in the entire system. Therefore, improving the performance of LID on short utterances is one of the important tasks in reducing system latency.

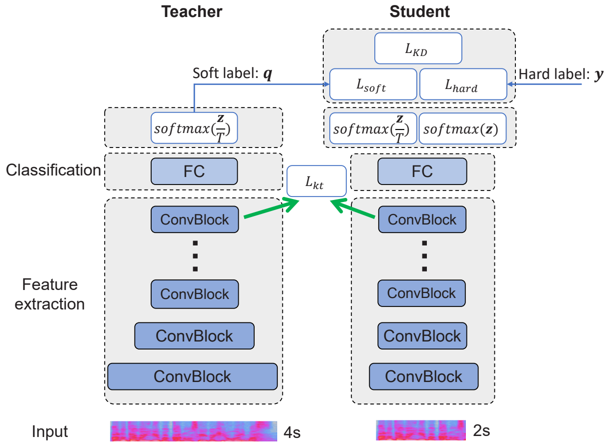


Fig. 2　The proposed KDRL method for short-utterance LID tasks

Compared to long utterances, the distribution of short utterances has a large intra-class variation, which results in large model confusion. Reducing this variation is expected to improve the performance for short-utterance LID tasks. Inspired by previous works of knowledge distillation [7], we proposed a knowledge distillation-based representation learning (KDRL) approach by transferring the representation knowledge of a long-utterance-based teacher model to a short-utterance-based student model [8].

The proposed KDRL method is illustrated in Fig. 2. Suppose Θ_t and Θ_s are the parameters of the neural networks providing the internal representation of a teacher network and a student network, respectively. The proposed KDRL is based on minimizing the following loss function:

$$\mathrm{L_{KDRL}} = \frac{1}{N}\sum_{\boldsymbol{x}_s,\boldsymbol{x}_t,\boldsymbol{y}}((1-\lambda)L_{hard}(\boldsymbol{x}_s,\boldsymbol{y}) \\ +\lambda L_{kt}(\boldsymbol{x}_t,\boldsymbol{x}_s,\Theta_t,\Theta_s)) \tag{1}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{x}_s$ are input samples of the teacher and student networks, respectively, and $\mathrm{L_{kt}}$ is a distance metric of the internal representation defined as

$$L_{kt}(\boldsymbol{x}_t,\boldsymbol{x}_s,\Theta_t,\Theta_s) = ||u_t(x_t;\Theta_t) - u_s(x_s;\,\Theta_s)|| \tag{2}$$

where $\|\circ\|$ is a norm function, for example L1- or L2-norm, which is used to measure the representation distance between the teacher and student models. The $\mathrm{u_t}$ and $\mathrm{u_s}$ are the teacher and student deep nested functions up to their respective selected layers with output of parameter set Θ_t and Θ_s, respectively.

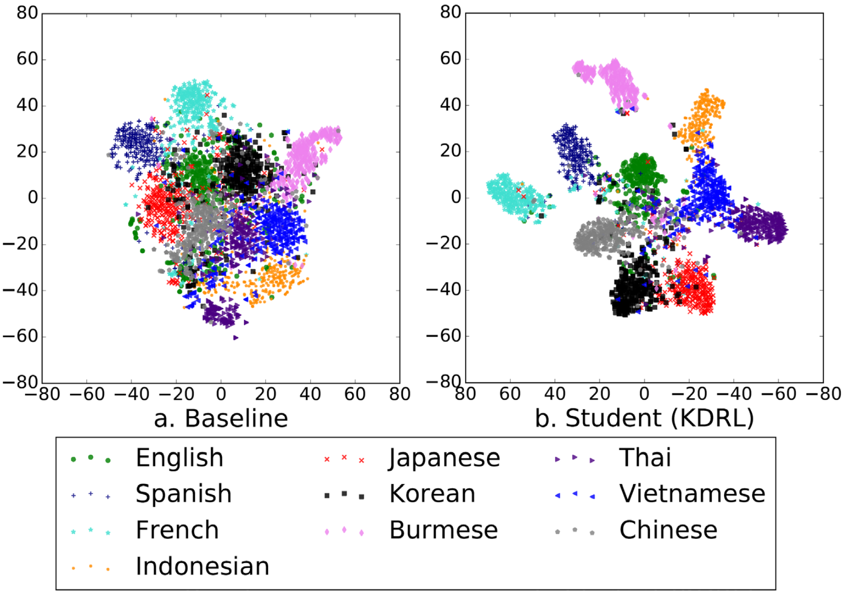
With the proposed method, the feature representation knowledge, corresponding to a hidden layer of a teacher model, is transferred to a student model to help the student model to capture robust discriminative information from short utterances. To understand the effect of the KDRL method, we plotted the distributions of the internal representations of the baseline and the KDRL method by using t-Distributed Stochastic Neighbor Embedding (TSNE) [9]. Fig. 3(a) is obtained from the deep convolutional neural network (DCNN) baseline model trained with 1-second utterances. Fig. 3(b) is obtained from the KDRL-based student model that was trained with 1-second utterances. During the student model training, a 4-second utterance-based teacher model was utilized. This figure shows that by reducing the internal representation difference between short and their corresponding long utterances, the student model could have a higher inter-class variation and lower intra-class variation than the baseline model.

We further investigated the KDRL method on the widely used language embedding technique, i.e., x-vector framework [6], and proposed an x-vector extraction approach with adding compensation constraint only for the mean component in the x-vector space. In the proposed vector, the mean component is expected to represent high-level abstract language information while retaining the variance component to encode frame-based local phonetic information for short utterances [10]. Another work was focused on reducing the difficulty of optimizing the student model with a fixed pre-trained teacher model because the inputs of the student model are short utterances while the inputs of the teacher model are the corresponding longer utterances. Such difference makes the student model easy to be stuck in a local minimum with a bad performance. In that work, rather than using a fixed pre-trained teacher model, we investigated an interactive teacher-student learning method to improve the optimization by adjusting the teacher model with reference to the performance of the student model [11].

2.2　Robustness for cross-domain/channel problem

The recent deep neural network-based LID technologies significantly improve the accuracy of LID by using a large amount of training data and complex network structure with powerful acoustic feature extraction and abstraction capability. However, in real applications, such techniques often suffer from overfitting problems because the recording conditions and speaking styles of a test dataset are different from those of the training dataset, i.e., the cross-domain problem.

To reduce the domain discrepancy, we proposed an optimal transport (OT)-based unsupervised neural adaptation framework for cross-domain LID tasks [12]. The OT is initial for finding an optimal transport plan to convert one probability distribution shape to another shape with the least effort [13]-[14]. In our work, we adopted the OT distance metric to measure the adaptation loss between source and target data samples. Let p_s and p_t are data distribution of the source and target domains, respectively. Then, the proposed adaptation methods can be described



a. Baseline　b. Student (KDRL)

Fig. 3　Representation distributions based on TSNE of the selected hidden layer on an NICT 10 language dataset

as,

$$\mathrm{L_T} = \min(L_{CE}^{S} + \lambda L_{OT}(p_s, p_t)) \tag{3}$$

where $\mathrm{L}_{\mathrm{CE}}^{S}$ is the multi-class cross-entropy classification loss in source domain, and $\mathrm{L_{OT}}(p_s, p_t)$ is the OT loss to measure the distribution distance between source and target data samples.

Based on our proposed unsupervised adaptation learning, we expect that the mismatch between training and testing will be reduced. Fig. 5 illustrates the effect of unsupervised adaptation on language cluster distributions. In this figure, only two languages are shown with label IDs as lang1 and lang5 from a testing data set. In Fig. 5(a), due to the difference in recording channels, the clusters belonging to the same language are separated (pairs of lang5 train vs. lang5 test and lang1 train vs. lang1 test). Since the classifier is designed based on the training data set, it is not strange that the performance of the baseline system on the testing data set is degraded. After adaptation, as shown in Fig. 5(b), the clusters of the testing data set are pushed to be overlapped with those of the training data set for the same language.

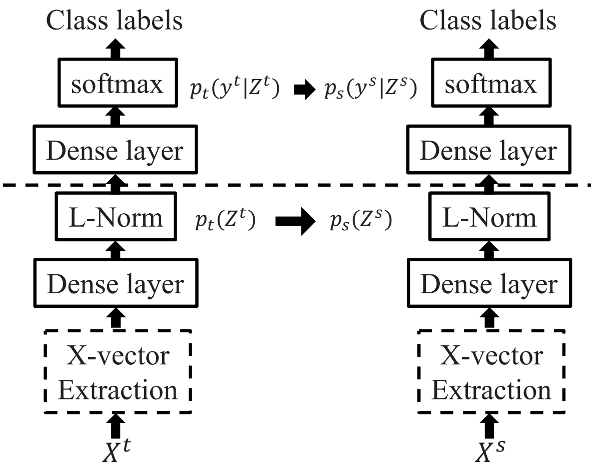
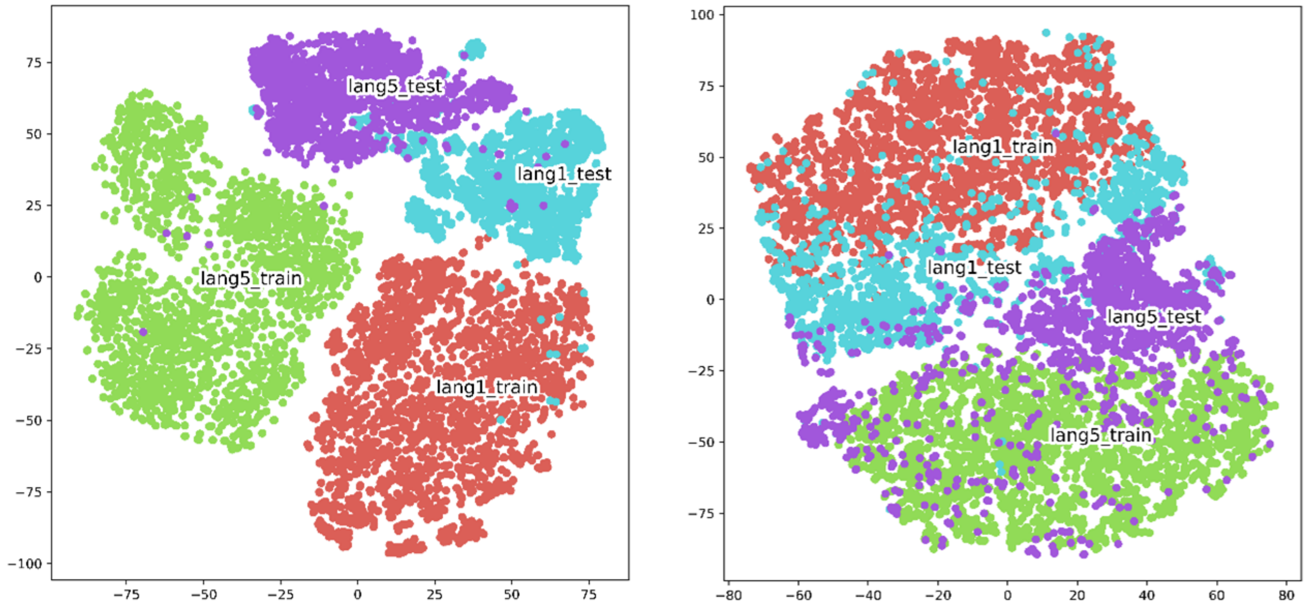


Fig. 4 The proposed unsupervised OT-based adaptation neural network for LID

Excepted the unsupervised approach, we also proposed to use linguistic features to improve the robustness of the LID tasks [15]. For LID tasks, not only acoustic features, such as phonotactics information, but also linguistic features, such as contextual information, are important cues to determine a language [1]. Therefore, we proposed a novel transducer-based language embedding approach by integrating an RNN transducer (RNN-T) model into a language embedding extraction framework that is illustrated in Fig. 6. Benefiting from the advantages of the RNN-T's linguistic representation capability and the proposed method can exploit both phonetically-aware acoustic features and explicit linguistic features for LID tasks. Our experimental results showed that compared with the conformer encoder-based baseline method, the proposed method obtained 38% and 28% relative improvement on in-domain and cross-domain datasets, respectively.



(a)

(b)

Fig. 5 Language cluster distributions based on the TSNE [9] for a test set in cross-domain language recognition task: before adaptation (a), and after adaptation (b)

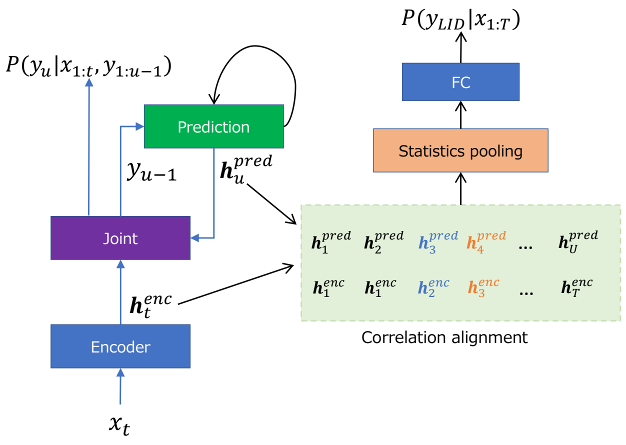

$P(y_{LID}|x_{1:T})$
FC
Statistics pooling
$P(y_u|x_{1:t}, y_{1:u-1})$
Prediction
y_{u-1}
h_u^{pred}
Joint
h_t^{enc}
Encoder
x_t
h_1^{pred} h_2^{pred} h_3^{pred} h_4^{pred} ... h_U^{pred}
h_1^{enc} h_1^{enc} h_2^{enc} h_3^{enc} ... h_T^{enc}
Correlation alignment

Fig. 6 The proposed RNN-transducer-based language embedding

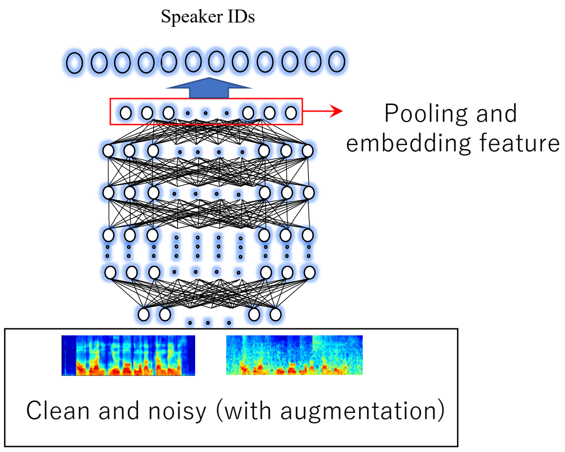

Speaker IDs
Pooling and embedding feature
Clean and noisy (with augmentation)

Fig. 7 Deep speaker embedding (x-vector) for speaker feature extraction

3 Speaker Recognition

Speaker information is useful for many real speech applications, for example, multi-speaker meetings, interviews, or dialogs, as well as speaker authentication for security access [16]. In most applications, besides recognizing the content of speech based on the ASR technique, speaker identity information should also be recognized based on the automatic speaker recognition technique. There are two basic tasks in speaker recognition: One is speaker verification (SV), and the other is speaker identification (SI). The conventional pipeline in constructing such speaker recognition system is composed of front-end speaker feature extraction and backend speaker classifier modeling. Front-end feature extraction tries to extract robust and discriminative features to represent speakers, and the backend classifier tries to model speakers with the extracted features for either classification or verification. Obviously, how to extract speaker representation is essential for achieving robust performance.

3.1 Deep speaker embedding for speaker recognition

One of the most representative features of front-end speaker extraction is the i-vector [17]. In i-vector extraction, speech utterances with variable durations can be converted to fixed dimension vectors with the help of Gaussian mixture models (GMM) on probability distributions of acoustic features. Due to the success of deep learning techniques in speech and image processing, several alternative speaker features have been proposed, e.g., d-vector [18] and x-vector [19]. In particular, x-vector is widely used as one of the speaker-embedding representations in most state-of-the-art frameworks [19]. The basic model architecture for speaker embedding is shown in Fig. 7. In this figure, different types of neural network architectures could be applied, e.g., dense-connected feedforward network (FFN), convolutional neural network (CNN), time delay neural network (TDNN), etc. The task in model training is for speaker recognition with speaker identity as target labels, and the pooling layer output is used as feature representation.

As illustrated in Fig.7, the advantage of x-vector representation is that the model for x-vector extraction could be efficiently trained with a large number of speech samples from various speakers. Moreover, in order to explore robust speaker information, data augmentation with various noise types and signal-to-noise ratios (SNRs) could be easily applied in model training [19]. The extracted speaker embedding feature could show excellent clustering properties. An example is shown in Fig. 8 (samples for 50 speakers are shown).

This cluster distribution is a feature projection based on the TSNE [9]. From this figure, we can see that speech samples are well separated based on their speaker identities. We believe that tasks related to speaker information based on this representation could obtain good performance.

3.2 Hybrid generative and discriminative backend modeling on speaker embedding

Based on speaker embedding (e.g., x-vector), various tasks could be constructed based on different back-ends related to specific tasks, for example, a probabilistic linear discriminant analysis (PLDA) [20] or joint Bayesian (JB) [21] modeling on the speaker embedding feature for various tasks as shown in Fig. 9.

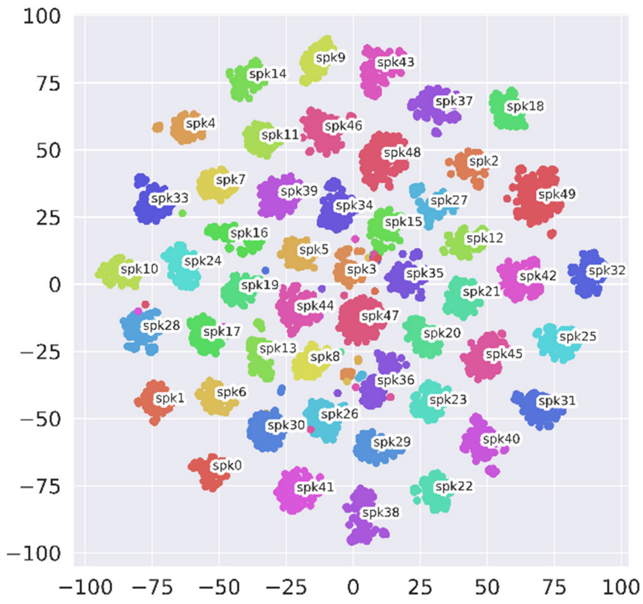


Fig. 8　Speaker clustering based on speaker embedding feature

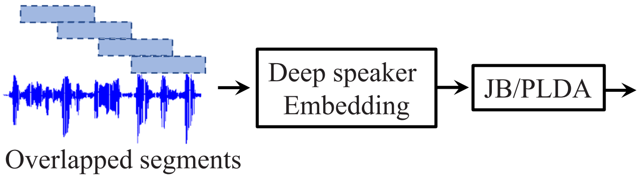


Fig. 9　Speaker embedding and backend modeling based on generative probabilistic models

In this figure, speech input is divided into several segments with a certain duration in speaker-embedding feature extraction. The extracted speaker embedding feature is modeled based on a probabilistic model. If the task is for SV, we need to estimate a log-likelihood ratio (LLR) for a hypothesis test as:

$$LLR = log\frac{p(X_i, X_j|H_S)}{p(X_i, X_j|H_D)} \tag{4}$$

where X_i, X_j are two compared feature vectors, H_S and H_D are hypotheses as the same speaker or different speakers. The generative probabilistic model is robust to various noise and unknown speakers but lacks in discriminative power. In our study, we proposed to integrate the generative model with a discriminative learning framework for improving the performance [22]. As the generative model focuses on class-conditional feature distributions while the discriminative model focuses on classification boundaries, the generative model could have a good generalization for short utterances (but less discriminative power), the discriminative model has a high discriminative capacity (but less generalization ability to short utterances). Fig. 10 shows the two different focuses of the two types of models. In this figure, only two classes are shown. By coupling the generative model in a discriminative neural network learning framework, we could combine both the advantages of generative and discriminative models to constrain large model variation.

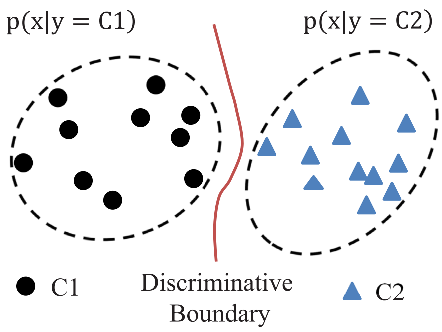


Fig. 10　Generative model (focuses on class conditional feature distributions indicated by dash-circles) vs. discriminative model (focuses on discriminative class boundary represented as solid curve). C1 and C2: class 1 and 2, respectively

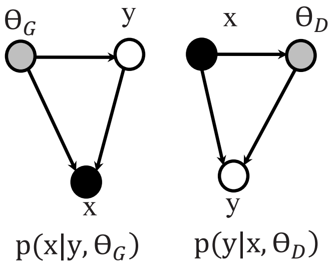


Fig. 11　Probabilistic graphic model for generative (left) and discriminative models (right)

Correspondingly, the probabilistic graphic network could be represented as in Fig.11. In this figure, **x** denotes a feature variable, y means a speaker ID label. In the generative model, the probability measures the likelihood with a given speaker ID label y to generate an acoustic observation **x**. In the discriminative model, the probability represents the posterior probability by given acoustic observation to estimate a speaker ID label. In Fig. 11, θ_G and θ_D are model parameter sets for the generative and discriminative models, respectively. In most studies, these two model parameter sets are estimated based on different methods. In our study, we proposed to couple the generative model with a discriminative learning framework in model parameter estimation. The proposed model framework is shown in Fig. 12.

In Fig. 12, the model framework was adopted for the SV task with two hypothesis labels H_S and H_D, i.e., the two compared utterances are from the same and different

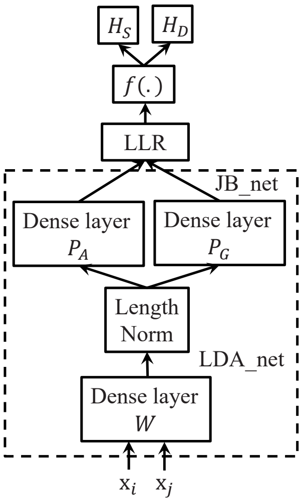


Fig. 12 The proposed two-branch Siamese neural network with coupling of the generative joint Bayesian model structure

speakers, respectively. LLR means log-likelihood ratio score calculated as in eq. 4, and JB_net as joint Bayesian model network (the generative model), LDA_net as a linear discriminative analysis network which was used for dimensional reduction. Dense layers were used to fit the functions of model parameters (coupling to the generative model) used in JB transform with parameter sets P_A and P_G (refer [22] for details). And the input feature vectors $\boldsymbol{x}_i, \boldsymbol{x}_j$ are two compared vectors representing two utterances. For discriminative training, we further proposed an objective function based on false alarm and miss measure metrics which are used in detection tasks. The idea was illustrated in Fig. 13. In this figure, by giving a decision threshold θ, the tradeoff between miss rate and false alarm rate could be estimated in optimization.

Based on the proposed framework, the two model distributions (H_S and H_D) were further separated as shown in Fig. 14. And the speaker verification experiments confirmed the improved performance.

4 Summary

In this paper, we gave an overview of our recent work on both spoken language identification and speaker recognition tasks. Our research work focused on both improving the basic theoretical methods and filling the gap between research and development so as to further promote the application of multilingual speech technology.

Recently, researchers have made a big progress on LID and SV/SI techniques. However, there are still some challenges that need to be overcome for these techniques to be applied well in real environments:

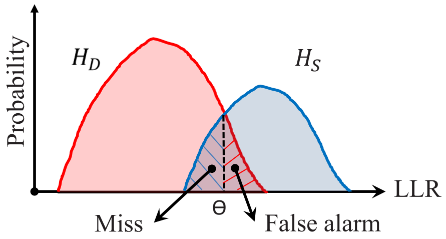


Fig. 13 The LLR distribution for H_S and H_D conditions with consideration of miss and false alarm

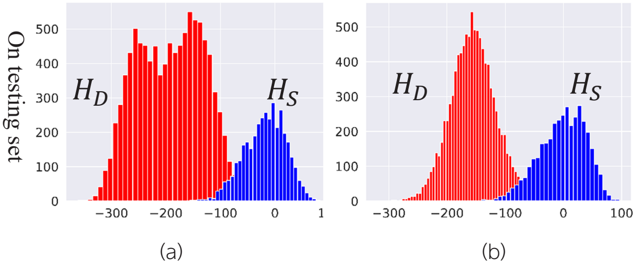


Fig. 14 LLR distributions for test data set in H_S and H_D spaces before (a) and after (b) the couple training. H_S: the same speaker hypothesis, H_D: the different speaker hypothesis

1. LID and SV/SI techniques often work as the preprocessing step of a speech processing system; therefore, the real-time factor (RTF) and latency are important factors. Recent state-of-the-art techniques are based on large self-supervised models, for example, wav2vec [23]; therefore, developing high-performance models with low RTF and latency is necessary, especially when the techniques are running on mobile devices.

2. Cross-domain/channel problem is still one of the most challenging problems for deep learning techniques. For LID tasks, it is more sensitive to the cross-channel problem, and the LID model may even extract channel features rather than language features in recognition tasks. In future work, we will continue focusing on improving the robustness of the LID and SV/SI by using self-supervised learning and pre-training techniques.

References

1 H. Li, B. Ma, and K. A. Lee, “Spoken language recognition: From fundamentals to practice,” Proc. The IEEE, vol.101, no.5, pp.1136–1159, 2013.

2 P. Shen, X. Lu, L. Liu, and H. Kawai, “Local Fisher discriminant analysis for spoken language identification,” Proc. ICASSP, 2016.

3 X. Lu, P. Shen, Y. Tsao, and H. Kawai, “Regularization of neural network model with distance metric learning for i-vector based spoken language identification,” Computer Speech & Language, vol.44, pp.48–60, 2017.

4 A. Lozano-Diez, R. Zazo Candil, J. G. Dominguez, D. T. Toledano, and J. G. Rodriguez, “An end-to-end approach to language identification in short utterances using convolutional neural networks,” Proc. of INTER-

SPEECH, pp.403–407, 2015.
5 S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional Modelling for Short Duration Language Identification," Proc. INTERSPEECH , pp.2809–2813, 2017.
6 D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," Proc. Odyssey, pp.105–111, 2018.
7 G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
8 P. Shen, X. Lu, S. Li, and H. Kawai, "Feature representation of short utterances based on knowledge distillation for spoken language identification." Proc. INTERSPEECH, 2018, pp.1813–1817.
9 L. Maaten and G. Hinton, "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research 9 (Nov.), pp.2579–2605, 2008.
10 P. Shen, X. Lu, K. Sugiura, S. Li, and H. Kawai, "Compensation on x-vector for Short Utterance Spoken Language Identification." Proc. Odyssey, 2020, pp.47–52.
11 P. Shen, X. Lu, S. Li, and H. Kawai, "Interactive learning of teacher-student model for short utterance spoken language identification." Proc. ICASSP, pp.5981–5985, 2019.
12 X. Lu, P. Shen, Y. Tsao, and H. Kawai, "Unsupervised Neural Adaptation Model Based on Optimal Transport for Spoken Language Identification," Proc. ICASSP, pp.7213–7217, 2021.
13 G. Peyre and M. Cuturi, "Computational Optimal Transport," ArXiv:1803.00567, 2018.
14 N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.36, no.9, pp.1853–1865, 2017.
15 P. Shen, X. Lu, and H. Kawai, "Transducer-based language embedding for spoken language identification," Proc. INTERSPEECH, 2022.
16 J. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," IEEE Signal processing magazine, vol.32, no.6, pp.74–99, 2015.
17 N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol.19, no.4, pp.788–798, 2011.
18 E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," Proc. ICASSP, pp.4052–4056, 2014.
19 D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," Proc. ICASSP, pp.5329–5333, 2018.
20 A. Sizov, K. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," Joint IAPR, International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp.464–475, 2014.
21 D. Chen, X. Cao, D. Wipf, F. Wen, and J. Sun, "An efficient joint formulation for Bayesian face verification," IEEE Transactions on pattern analysis and machine intelligence, vol.39, pp.32–46, 2016.
22 X. Lu, P. Shen, Y. Tsao, and H. Kawai, "Coupling a Generative Model With a Discriminative Learning Framework for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, vol.29, pp.3631–3641, 2021.
23 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, 2020.

沈 鵬 （しん ほう）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声認識、言語識別、話者識別、イベント検出

Xugang Lu
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声認識、機械学習