様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科研費 KAKENHI

科学研究費助成事業　　研究成果報告書

令和　４ 年　６ 月 １６ 日現在

機関番号：８２６３６
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０３５

研究課題名（和文）Construction of a computational model to deal with the cocktail-party problem for intelligent speech interface

研究課題名（英文）Construction of a computational model to deal with the cocktail-party problem for intelligent speech interface

研究代表者

ＬＵ　Ｘｕｇａｎｇ（Lu, Xugang）

国立研究開発法人情報通信研究機構・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・主任研究員

研究者番号：２０３６２０２２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：カクテルパーティーのシナリオでは、さまざまな音源ソースを識別するため、多くの情報を分析することが必要です。本研究では、次のような成果を得られました。1.発話元を特定するために、誰が話しているかは最も重要な情報の1つです。話者埋め込みシステムの開発に加えて、話者認識のための生成的学習と識別的学習の結合を提案しました。私たちのフレームワークは、最先端のモデルと比較して大幅な改善を示しました。2.音声ソースの録音環境はドメインごとに変わる可能性があるため、教師なしドメイン適応手法の新しい距離メトリックを提案しました。提案されたアルゴリズムを応用し、クロスドメイン認識タスクで大幅な改善が得られました。

研究成果の学術的意義や社会的意義
カクテルパーティーのシナリオでは、混合音声ソースの場合、誰が話し、どの言語が使用されているかが、音声ソースの分離に関する重要な事前知識です。話者の認識性能と言語認識を改善するための新しいアイデアとアルゴリズムを開発しました。これは、音声ソースの事前知識の質を高めるのに役立ちます。

研究成果の概要（英文）：In cocktail party scenarios, many information need to be explored in order to identify different speech (or sound) sources. Under this project, we have the following contributions: 1. For identifying speech source, who is speaking (speaker information) is one of the most important information. Besides developing speaker embedding system, we proposed a coupling of generative and discriminative learning for speaker recognition. Our framework showed a large improvement compared with state of the art models. 2. Concerning speech source recording environments may change in different domains, we proposed a new distance metric for unsupervised domain adaptation technique. We have tested the proposed adaptation algorithm on both speaker and language recognition tasks, and obtained promising improvement when speech recording environments are changed.

研究分野：人工知能、信号処理

キーワード：知能情報

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

For most speech technology application systems, when they are applied in cocktail-party scenarios, i.e., real acoustic environments with mixed sound sources, the performance is drastically degraded. The reason is that the conventional computational models (no matter with or without deep learning algorithms) used in those applications take all entangled sound sources without actively selecting one of (or some of) them for processing and recognition. Parsing mixed sound sources in cocktail-party scenarios is a very important "intelligent" function of human beings during speech communication. In this study, our ambition is to construct a new computational model to realize this "intelligent" function with selective attention for speech interface. The computational model should integrate both bottom-up sound saliency detection and top-down selective attention processing to dynamically parse the incoming mixed sound sources. Among this computational framework, integrating prior knowledge is one of the most effective ways for parsing the mixed speech sources. Among the available prior information, speaker information and language information are important factors that could be accurately identified before speech source separation.

２．研究の目的

The speaker and language identification algorithms could be directly applied for our purpose. However, there are several problems remained in real applications: 1. For short utterance, it is difficult to obtain a satisfied performance due to limited information (speaker information for speaker recognition task, and language information for language recognition task). 2. For a model trained with a certain domain data, the performance will degrade drastically when the domain is different in testing (e.g., real recording environments are changed frequently). The purpose of this study is to deal with the two problems for real application purpose.

３．研究の方法

For the first problem we mentioned above, besides adopting an embedding method (speaker or language embedding), we further propose to integrate a generative model and a discriminative model for improving the performance. 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 high discriminative capacity (but less generalization ability to short utterances). Fig. 1 shows the two different focuses of the two types of models. In this figure, only two classes are showed. By coupling 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 (due to large feature variation of short utterances).

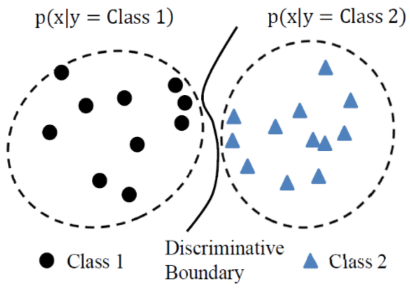


Fig. 1. Generative model learning focuses on class conditional feature distributions (dashed-circles of feature distribution shapes), and discriminative model learning emphasizes the class discriminative boundary (solid curve).

For the second problem, we adopt an adaptation method to deal with the cross-domain problem. The problem can be well described as: source domain data set $D_s = \{(x_i^s, y_i^s)\}_{i=1,2,\dots,N}$, and target domain data set $D_t = \{(x_i^t, y_i^t)\}_{i=1,2,\dots,M}$. Due to domain changes (e.g. recording channels), there exists domain discrepancy, i.e., $p_s(x,y) \neq p_s(x,y)$. The purpose for domain adaptation is to reduce this discrepancy to make $p_s(x,y) \approx p_s(x,y)$. Based on Bayesian theory, $p_s(x,y) = p_s(y|x)p_s(x)$ and $p_t(x,y) = p_t(y|x)p_t(x)$, we need to approximate the two terms as: $p_s(y|x) \approx p_t(y|x)$ and $p_s(x) \approx p_t(x)$. For finding a latent transformed space $z = \varphi(x)$, the approximation will be $p_s(y|z) \approx p_t(y|z)$ and $p_s(z) \approx p_t(z)$.

４．研究成果

For dealing with the first problem, we proposed coupling a generative model with a discriminative learning framework, and applied to improve speaker recognition performance. The proposed model framework is showed in Fig. 2.

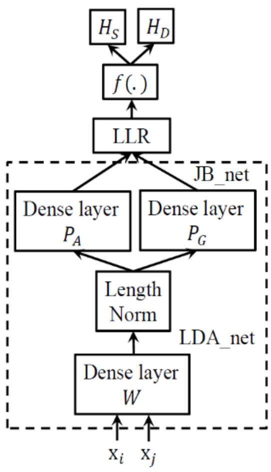


Fig. 2 The proposed two-branch Siamese neural network with coupling of the generative joint Bayesian model structure.

In this Fig. 2, the model framework was adopted for speaker verification task with two hypothesis labels H_S and H_D as the two compared utterances are from the same speaker and different speakers, respectively. LLR means log-likelihood ratio score, and JB_net as joint Bayesian model (as generative model), LDA_net as linear discriminative net for dimensional reduction. Dense layers were used to fit the functions of model parameters (coupling to the generative model). And the input feature vectors x_i and 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. 3.

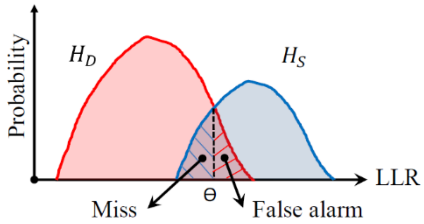


Fig. 3 The LLR distribution for H_S and H_D conditions.

Based on the proposed framework, the two hypothesis distributions (H_S and H_D) were further separated as showed in Fig. 4. And the speaker verification experiments confirmed the improved performance.

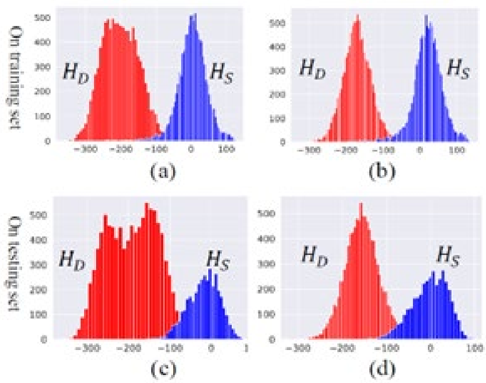


Fig 4. LLR distributions in H_S and H_D spaces: the first row (a, b) and second row (c, d) are for the training and test sets, respectively; the left column (a, c) for setting with generative model parameters learned based on the EM algorithm, and the right column (b, d) for setting with discriminatively trained parameters after initializing with generative model parameters.

For dealing with the cross-domain problem, our proposed model framework is showed in Fig. 5. In this framework, x^t and x^s represent feature vectors from target and source domains, with a transform (X-vector extraction and a neural dense layer transform, and length normalization), we obtained latent feature representations as z^t and z^s, and finally obtained their classification labels y^t and y^s. Based on our theoretical analysis, we optimized the model parameters for approximation of joint distributions for cross-domain adaptation.

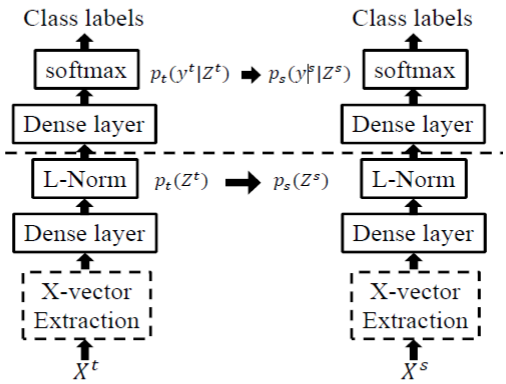
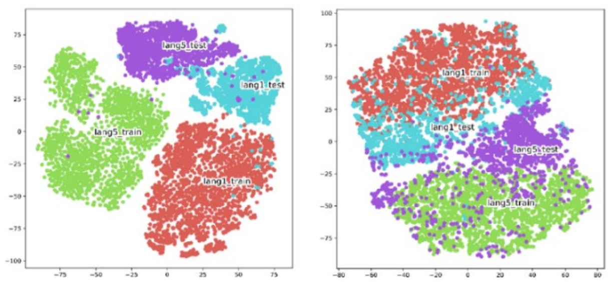


Fig. 5 Joint distribution adaptation for language recognition.

Based on the proposed adaptation algorithm, the domain discrepancy was reduced. An example of the feature distributions of two languages before and after adaptation was showed in Fig. 6. From this figure, we can see that after adaptation, the distributions of training and test sets have large overlap.



(a) (b)

Fig. 6 Language cluster distributions based on the TSNE for cross-domain test: before adaptation (a), and after adaptation (b).

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Xugang Lu, Peng Shen, Sheng Li, Yu Tsao, Hisashi Kawai	４．巻 29
２．論文標題 Coupling a Generative Model With a Discriminative Learning Framework for Speaker Verification	５．発行年 2021年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 3631-3641
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2021.3129360	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Xugang Lu, Peng Shen, Yu Tsao, Hisashi Kawai
２．発表標題 Class-Wise Centroid Distance Metric Learning for Acoustic Event Detection
３．学会等名 Interspeech 2019（国際学会）
４．発表年 2019年

１．発表者名 Peng Shen, Xugang Lu, Komei Sugiura, Sheng Li, Hisashi Kawai
２．発表標題 Compensation on x-vector for short utterance spoken language identification
３．学会等名 Odyssey 2020 The Speaker and Language Recognition Workshop（国際学会）
４．発表年 2020年

１．発表者名 Xugang Lu, Peng Shen, Yu Tsao, Hisashi Kawai
２．発表標題 UNSUPERVISED NEURAL ADAPTATION MODEL BASED ON OPTIMAL TRANSPORT FOR SPOKEN LANGUAGE IDENTIFICATION
３．学会等名 ICASSP2021（国際学会）
４．発表年 2021年

１．発表者名 Xugang Lu, Peng Shen, Sheng Li, Yu Tsao, Hisashi Kawai
２．発表標題 Siamese Neural Network with Joint Bayesian Model Structure for Speaker Verification
３．学会等名 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関