UNSUPERVISED CHANNEL ADAPTATION FOR LANGUAGE IDENTIFICATION USING CO-TRAINING

Sriram Ganapathy, Mohamed Omar, Jason Pelecanos

IBM T.J Watson Research Center, Yorktown Heights, NY, USA.

ABSTRACT

Language identification (LID) of speech signals in conditions like adverse radio communication channel is a challenging problem. In this paper, we address the scenario of improving the performance of a LID system on mis-matched radio communication channels (not seen in training) given a small amount of speech data without language labels. We develop a co-training procedure using two diverse acoustic LID systems to improve the performance by effectively utilizing the adaptation data. The acoustic LID systems use different features, projection methods and back-end classifiers. Assuming that the classification errors for the diverse LID systems are independent, the co-training procedure improves the classification accuracy of each system. Various LID experiments are performed on the mis-matched channels in a leave-one-out setting for a variety of noise conditions. In these experiments, with small amounts of unsupervised data from the new channel, we show that the proposed co-training procedure provides significant improvement (average relative improvement of 32 %) over the baseline scenario of no-adaptation and noticeable improvements of about 10 % over a self-training framework.

Index Terms— Radio Channel Speech, Language Identification, Co-training, Unsupervised Adaptation.

1. INTRODUCTION

The speech signal received from a typical radio communication channel has artifacts which are different from additive noise or convolutive distortions like reverberations. The signal degradation in this scenario includes linear frequency transpositions, non-linear amplitude scale variation over a long-time span and harmonic distortions [1]. The DARPA program named robust automatic transcription of speech (RATS) targets the development of speech systems operating on highly distorted speech recorded over “degraded” radio channels. The data consists of recordings obtained from retransmitting a clean signal over eight different radio channel types, where each channel introduces a unique degradation mode specific to the device and modulation characteristics [1].

Recently, the language identification task was performed on this data using the same channels in training and testing (matched conditions) [2, 3]. Although reasonable language identification performance is obtained in this case, the set of eight channels does not represent the realistic scenario in which various other radio communication network and device characteristics can cause severe mis-match with the training conditions. In order to simulate these effects, we consider the leave-one-out setting where one of the eight channels is not used in training (mis-matched conditions). Since specific device operating points, modulation type, carrier bandwidths and transmission path artifacts influence the acoustic signal, the performance is severely degraded in these mis-matched conditions.

In this paper, we consider the problem of enhancing the LID system performance given a small amount of unsupervised adaptation data from the new channel. Recently, supervised adaptation with small amounts of data was shown to improve the performance for a new channel [4]. In the past, the use of unsupervised data to improve the performance on mis-matched telephone channels was studied for speaker recognition using background model synthesis [5] and feature mapping [6]. However, the artifacts introduced by a radio communication channel are more non-linear and time varying compared to the linear convolutive effects seen in telephone channels.

In this paper, we address the problem of unsupervised channel adaptation using the co-training algorithm [7]. Co-training is a learning procedure which utilizes the independence among multiple diverse weak classifiers. The most confident examples of one classifier are used to retrain the other classifier and vice-versa. It has been shown that co-training can provide considerable improvements using moderate assumptions of conditional independence among the classifiers [8]. In the past, co-training was successfully applied to various tasks like email classification [9], dialect identification [10], speech summarization [11] and gesture recognition [12].

For the co-training of the LID systems in an unsupervised channel adaptation setting, we develop two diverse acoustic systems. The two systems use different front-end representations, projection models as well as back-end classifiers. Various LID experiments are performed with small amount of adaptation data from the new channel. In these experiments, we show that the adaptation procedure using the proposed co-training framework provides significant improvements over the baseline unseen channel setting (average relative improvements of about 32%) as well as a self-training scenario (average relative improvements of about 10%). We also show that the proposed co-training procedure can be used in conjuction with system combination approaches which are typically used in language recognition systems (for example, [3]).

The rest of the paper is organized as follows. In Sec. 2, we describe the co-training framework of semi-supervised learning. A brief description of the LID systems is given in Sec. 3. Adaptation experiments in the unsupervised setting are described in Sec. 4. We also report additional experiments with varying amounts of adaptation data. In Sec. 5, we conclude with a summary of the paper.

2. SEMI-SUPERVISED LEARNING USING CO-TRAINING

Co-training is a semi-supervised learning algorithm where multiple weak classifiers (trained with supervised data) are used together to boost the learning from the unsupervised data. The algorithm works by using the most confidently classified examples from one classifier

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

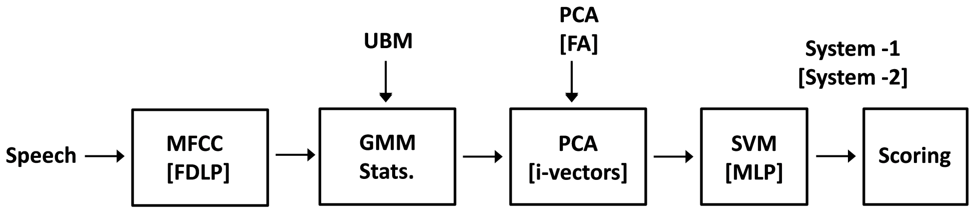

Fig. 2. Block schematic of LID systems used in co-training.

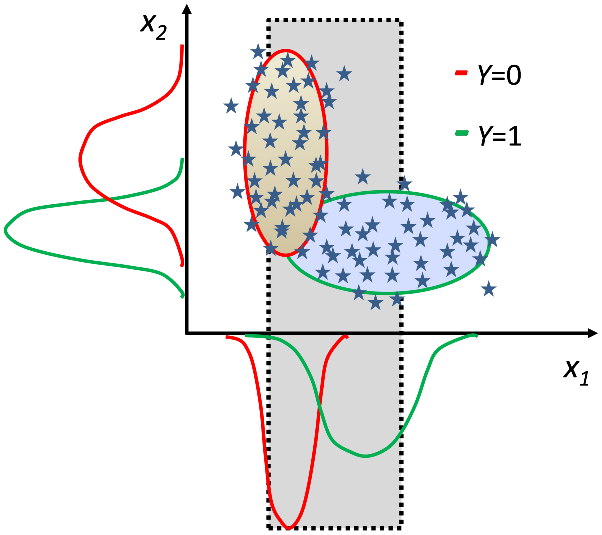

Fig. 1. Illustration of co-training using a two dimensional Gaussian example. Here, the confident labels from classifier $h_1(x_1)$ (identified as points lying outside the shaded region) are used to re-train the classifier $h_2(x_2)$.

to improve the learning of the other classifier. The learning algorithm depends on the conditional independence of the classifiers given the class labels [7]. The performance improvements from co-training are significant when this assumption is validated [8].

We illustrate the co-training algorithm using a simple example [13]. Consider a binary classification problem on the feature space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1$,$\mathcal{X}_2$ correspond to two different views of the data. For a given data sample $x = [x_1, x_2]$, the assumption of class conditional independence is the following,

$$\mathcal{P}(x_1, x_2|y) = \mathcal{P}(x_1|y)\mathcal{P}(x_2|y) \;\; for \;\; y \; \in \; \{0, 1\}. \tag{1}$$

The algorithm works in the following way. A weak classifier, $\mathrm{h}_1(x_1)$, trained using the labeled data, is applied to the unlabeled data. The examples with high confidence are selected and are used along with the labeled data to train classifier $\mathrm{h}_2(x_2)$ on the second view x_2. This process is repeated with roles of h_1 and h_2 reversed.

The intuition behind the algorithm is depicted using a two dimensional example in Fig. 1. In this figure, the views x_1 and x_2 are assumed to be one dimensional and the class conditional distributions are assumed to be jointly Gaussian and uncorrelated. The points scattered in the plot represent unlabeled data. The data points inside the marked region are ignored and the points outside represent the confident examples selected by the classifier $\mathrm{h}_1(x_1)$. As seen in this plot, the examples chosen for re-learning the classifier $\mathrm{h}_2(x_2)$ represent unbiased training samples over the range of x_2 due to the conditional independence assumption. This can boost the learning of the classifier h_2 and a similar procedure can be repeated by reversing the roles of h_1 and h_2.

3. LID SYSTEMS FOR CO-TRAINING

To adapt the LID systems to noisy acoustic conditions induced by a new radio communication channel, we use two different views generated by LID systems which use diverse acoustic features, projection methods as well as back-end classifiers. The block schematic of the two systems is shown in Fig. 2.

3.1. MFCC-PCA-SVM System

The input signal is processed using Wiener filtering [14] and Mel-frequency cepstral coefficients [15] are derived from 37 Mel-bands in the $125 - 3800$ Hz range. We derive 14 cepstral features which are used to estimate shifted-delta-coefficients (SDC) [16] (with a 7-1-3-7 configuration). A Gaussian mixture model-universal background model (GMM-UBM) with 1024 components is trained using the training and development portion of the LID data [1]. The adapted Gaussian mixture means are concatenated to form the supervector (SV). We train a principal component analysis (PCA) projection model with 800 dimensions on the SVs. The reduced dimension PCA vectors are used in training support vector machines (SVM) for each language of interest with a third order polynomial kernel [2].

3.2. FDLP-FA-MLP System

Frequency domain linear prediction (FDLP) represents an autoregressive modeling technique for deriving the sub-band Hilbert envelopes [17]. These sub-band envelopes represent temporal modulation information in each sub-band. The FDLP envelopes are integrated in short-term windows (32 ms with a shift of 10 ms) to derive cepstral coefficients which are used to construct 98 dimensional SDC features. We train a GMM-UBM model using the FDLP features with 1024 components. The zeroth and first order GMM statistics for each recording are obtained and these are used for training a factor analysis (FA) model [18]. We use 300 dimensional i-vectors derived from the FA model to train a three layer multi-layer perceptron (MLP). The MLP is trained with 500 hidden units and uses a soft-max function at the output nodes. We use the standard back propagation learning with cross entropy error function.

3.3. Learning from Unsupervised Data

Since the projection methods used in the two systems are unsupervised, the unsupervised data from the channel of interest can be used

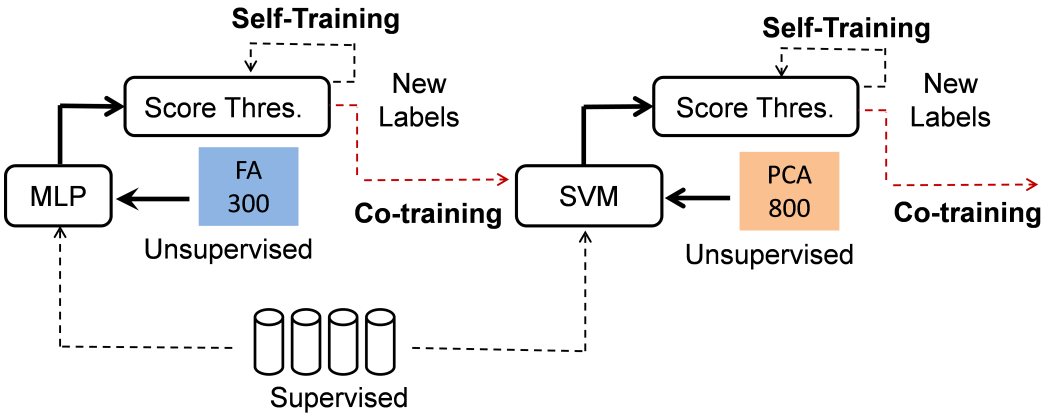

Fig. 3. Unsupervised adaptation with co-training using labels generated from the FDLP-FA-MLP system to re-learn the MFCC-PCA-SVM system. Here, we also show the self-training alternative to learning from unsupervised data.

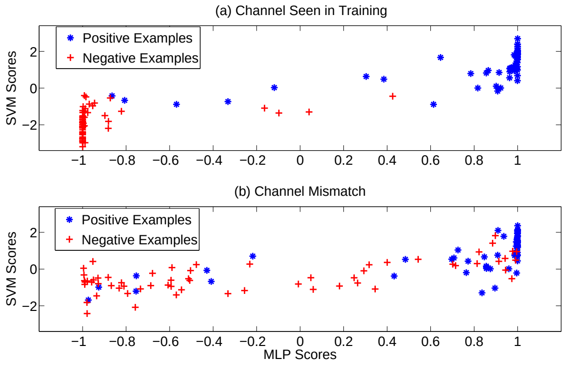

Fig. 4. Scatter plot of the FDLP-FA-MLP versus MFCC-PCA-SVM system scores for data from channel-H under two conditions ; (a) channel-H used in training and (b) channel-H not used in training.

in addition to the original data for retraining the projection models (either FA or PCA). The architecture used for classifier re-training from the unsupervised adaptation is shown in Fig. 3. Here, we contrast the traditional method of self-training using the unsupervised samples with the co-training framework.

In the past, co- training algorithm was used where the amount of unlabeled data was much larger than the amount of the labelled data. On the contrast, this paper explores the scenario where the unlabeled data is scarce. As described in Sec. 2, the co-training procedure requires the validation of the conditional independence assumption of the two systems. In the case of LID systems, this represents the scores obtained for each system on the unseen data. Fig. 4 plots the scores of the FDLP-FA-MLP system versus the MFCC-PCA-SVM system for one language ($URDU$). We plot the scores for two scenarios, (a) The first one represents the condition where the channel of interest (in this case, channel-H) is seen is training and (b) the second scenario where the channel of interest is not seen in training. As seen in this plot, the scores from the two systems are highly correlated when the channel is seen in training. In the mis-matched channel case, the scores from the two systems are less correlated (which would mean more independence for the joint Gaussian case) and the confident examples for one system represent informative examples for re-training the other system. In short, the assumptions for co-training are validated better in a mis-matched channel case.

4. EXPERIMENTS

The development and test data for the LID experiments use the LDC releases of the Phase-I RATS LID development [1]. This consists of speech recordings from previous NIST-LRE clean recordings as well as other RATS clean recordings passed through eight (A-H) noisy communication channels. The five target languages are Arabic, Farsi, Dari, Pashto and Urdu. In addition to this, the database consists of several other imposter languages. In our experiments, the UBM is trained using $39,123$ recordings from the matched channels and the FA/PCA models are trained with $65,078$ recordings. Separate UBM and projection models are trained for each leave-one-out setting. The test set consists of $14,328$ recordings from the eight noisy channels. The recordings used here contain about 120 seconds of speech. The training data contains about 270 hours per channel.

We compare the performance of the FDLP-FA-MLP system (denoted as MLP) and the MFCC-PCA-SVM system (SVM). Since it is typical in state-of-the-art language recognition systems to perform system combination using linear fusion, we also report the performance using a linear combination with equal weighting ($COMB$). The choice of equal weighting avoids the requirement of validation data from the mis-matched channels in determining the combination weights.

In the first set of experiments (Table 1), we use channel-D as the mis-matched channel with 2 hours of data from each language (12 hours in total) used for unsupervised adaptation. We report the average performance on the matched channels as well as the performance on the mis-matched channel (channel-D in this case). For self-training as well as co-training techniques, we use only one iteration by choosing one third of the confident examples (4 hours of development data). More iterations were not used as the amount of adaptation data was small compared to typical co-training applications with large amounts of unsupervised data.

As seen in Table 1, the performance is severely degraded when the channel of interest (in this case channel-D) is not used in training. The performance is improved by incorporating the unsupervised data

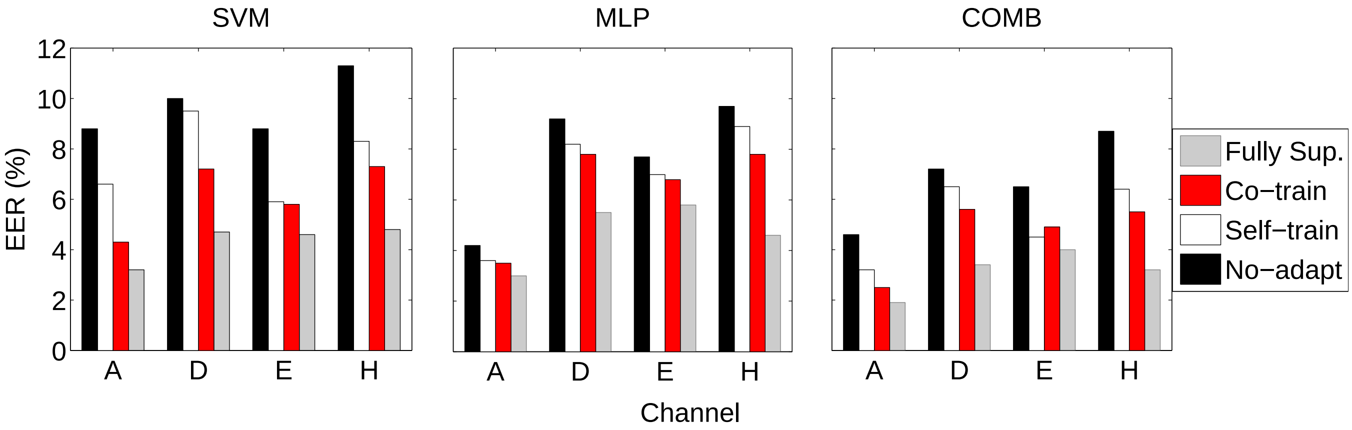

Fig. 5. Performance (EER %) of the LID system using 12 hours of unsupervised data for various leave-one out channels (A,D,E,H) with the baseline mis-matched channel for classification and subspace re-training (no-adapt), unsupervised adaptation with self-training, co-training and supervised adaptation.

Table 1. Performance (EER %) of the LID system using 12 hours of unsupervised development data from channel-D using MLP, SVM and COMB systems for seen channels (average performance over seven channels) and mis-matched channel-D.

Cond.	SVM		MLP		COMB	
	Seen	Mis-mat.	Seen	Mis-mat.	Seen	Mis-mat.
Completely Unseen	1.6	11.3	2.1	11.2	1.4	9.4
Projection Retrain	1.6	10.0	2.2	9.2	1.4	7.2
Self Training	1.5	9.5	2.3	8.2	1.3	6.5
Co-training	1.4	7.2	2.4	7.8	1.3	5.6
Supervised Adaptation	1.3	4.7	2.3	5.5	1.3	3.4

in retraining the projection models. This is consistent for re-learning the PCA as well as FA models. The self-training framework used in the MLP/SVM models provides improvements of about 5 % relative for SVM models and about 10 % relative for the MLP and COMB models. The co-training provides significant improvements over the self-training framework for the SVM system and the COMB system (with relative improvements of 25 % in the SVM system and 14 % in the COMB system). Compared to the scenario without any adaptation, the co-training framework improves the performance relatively by about 22% which amounts to reducing 42 % of the gap between supervised and unsupervised adaptation scenarios.

We repeat the unsupervised adaptation experiments using the leave-one-out strategy for a variety of channels (A,D,E and H). These channels represent a variety of modulation and noise characteristics seen in radio communication networks [1]. For example, channel A represents a narrow-band FM modulation (NFM) with carrier offset at the receiver, channel-D represents single side-band modulation with linear frequency shift and channel-H represents a NFM transmitter with a wide-band FM receiver.

In these experiments, we use 12 hours of unsupervised adaptation data which is used for re-training the PCA/FA subspaces. The performance for the mis-matched channel alone is reported for the SVM, MLP and COMB systems used in no-adaptation (projection re-training), self-training and co-training and supervised mode (shown in Fig. 5). The supervised adaptation mode represents the oracle performance and forms the upper bound for the unsupervised methods. On average, the co-training procedure provides a 19 % relative improvement over the self-training procedure for the SVM system and 7 % relative improvement for the MLP system. On the COMB system, the co-training provides 10 % relative improvement compared to the self-training system in terms of overall EER and about 26 % improvement over the self-training compared to the supervised adaptation upper-bound. Furthermore, the proposed adaptation framework improves the baseline scenario of no-adaptation by about 32 % using only a small amount of unsupervised adaptation data.

5. SUMMARY

In this paper, we have explored the application of the co-training learning algorithm for unsupervised adaptation of LID systems to speech data from a new radio channel. We use diverse acoustic LID systems based on MFCC/FDLP features, with different projections schemes (PCA/FA) and back-end classifiers (SVM/MLP). The diversity of the systems enhances the co-training learning technique where the unsupervised data is used with one system to generate confident labels for re-learning the other system. Various experiments performed using a small amount of adaptation data from a new channel show that the proposed co-training provided significant improvements over the baseline. In future, we plan to investigate this framework for the scenario with large amounts on unsupervised adaptation data, the combination of the co-training approach with robust speech features and with multiple LID systems.

6. ACKNOWLEGMENTS

The authors would like to thank Sibel Yaman, Kyu Han and Weizhong Zhu for their software contributions in implementing the RATS LID systems and Salim Roukos for discussions.

7. REFERENCES

[1] Walker K. and Strassel S., "The RATS Radio traffic collection system," in *Odyssey Speaker and Language Recognition Workshop*. ISCA.

[2] S. Yaman, J. Pelecanos, and M. Omar, "On the use of non-linear polynomial kernel SVMs in language recognition," in *Proceedings of Interspeech*, 2012.

[3] P. Matejka and et al., "Patrol team language identification system for DARPA rats p1 evaluation," in *Proceeding of Interspeech*, 2012.

[4] S. Ganapathy, M. Omar, and J. Pelecanos, "Noisy channel adaptation in language identification," in *IEEE Workshop on Spoken Language Technology*, 2012.

[5] L.P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97*, vol. 2, pp. 1071–1074.

[6] D.A. Reynolds, "Channel robust speaker verification via feature mapping," in *Acoustics, Speech, and Signal Processing, 2003. ICASSP-03*, vol. 2, pp. II–53.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[8] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*. ACM, 2000, pp. 86–93.

[9] S. Kiritchenko and S. Matwin, "Email classification with co-training," in *Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research*. IBM Press, 2001, p. 8.

[10] Gu Mingliang, Xia Yuguo, and Yang Yiming, "Semi-supervised learning based chinese dialect identification," in *International Conference on Signal Processing 2008*. IEEE, 2008, pp. 1608–1611.

[11] S. Xie, H. Lin, and Y. Liu, "Semi-supervised extractive speech summarization via co-training algorithm," in *Proc. of Interspeech*, 2010.

[12] C.M. Christoudias, R. Urtasun, A. Kapoorz, and T. Darrell, "Co-training with noisy perceptual observations," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2844–2851.

[13] H. Veeramachaneni, "The redundance of view-redundancy for co-training," in *http://mlstat.wordpress.com/tag/class-conditional-independence/*.

[14] A. Adami and et al., "Qualcomm-ICSI-OGI features for ASR," in *Seventh International Conference on Spoken Language Processing*, 2002.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[16] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and JR Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Seventh International Conference on Spoken Language Processing*, 2002.

[17] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *Signal Processing Letters, IEEE*, vol. 15, pp. 681–684, 2008.

[18] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.