Improved DNN-HMM English Acoustic Model Specially For Phonotactic Language Recognition

Wei-Wei Liu, Guo-Chun Li, Cun-Xue Zhang, Hai-Feng Yan, Jing He, Ying-Xin Gan, Yan-Miao Song,
Jian-Hua Zhou, Jian-Zhong Liu, Ying Yin, Ya-Nan Li*, Yu-Bin Huang, Ting Ruan, Wei Liu,
Rui-Li Du, Hua-ying Bai, Wei Li, Sheng-Ge Zhang
Department of Electronic Engineering, Tsinghua University, Beijing 100842, China, liu-ww10@hotmail.com
**Academy of Military Science, Beijing 100091, China, bennyhappy@gmail.com*

***Abstract*—The now-acknowledged sensitive of Phonotactic Language Recognition (PLR) to the performance of the phone recognizer front-end have spawned interests to develop many methods to improve it. In this paper, improved Deep Neural Networks Hidden Markov Model (DNN-HMM) English acoustic model front-end specially for phonotactic language recognition is proposed, and series of methods like dictionary merging, phoneme splitting, phoneme clustering, state clustering and DNN-HMM acoustic modeling (DPPSD) are introduced to balance the generalization and the accusation of the speech tokenizing processing in PLR. Experiments are carried out on the database of National Institute of Standards and Technology language recognition evaluation 2009 (NIST LRE 2009). It is showed that the DPPSD English acoustic model based phonotactic language recognition system yields 2.09%, 6.60%, 19.72% for 30s, 10s, 3s in equal error rate (EER) by applying the state-of-the-art techniques, which outperforms the language recognition results on both TIMIT and CMU dictionary and other phoneme clustering methods.**

***Keywords*-DPPSD English acoustic model; Phonotactic language recognition**

I. Introduction

Language recognition is the process of identifying a language from an utterance, which is an enabling technology in many applications, such as spoken document retrieval, speech translation, information security and forensics and multilingual speech recognition [1]. Currently, acoustic systems [2] and phonotactic systems [1] are two broad kinds of language recognition systems that are widely used. Phonotactic language recognition (PLR) is based on the assumption that phonotactic constraints contain information to identify the languages.

A typical parallel phone recognizer followed by vector space model (PPR-VSM) language recognition system is illustrated in Figure 1, where a collection of parallel phone recognizers is employed to convert the utterances into phone lattices by Viterbi algorithm according to the given acoustic model of phonemes without language models, then the lattices are used to perform phonotactic analysis to classify languages in Support Vector Machine (SVM) [3]. It is obvious that the performance of the phone recognizer front-end affects the succeeding proceeding work of the language recognition system. So the work of building a dramatic acoustic model of phone recognizer plays an important role in language recognition system.

Usually, each phone recognizer has a phone inventory of a single specific language. To deduce the effect of the incorrectness in phone tokenizing, the phone recognizers in the PPR-VSM system are usually trained either on multiple features like Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP) or on different acoustic models [4], such as GMM-HMM, ANN-HMM [5] and DNN-HMM and then fuse the recognition results. All these methods achieve good performance but need to repeat the training process as much as the number of front-ends and the type of feature adopted, which usually compute at high computational cost.

Researchers also find that in training material, native speakers do not actually speak in the way the language is written or listening materials is represented. The perceptual saliency of spoken English is often reduced creating variation in the way English is spoken in very formal contexts from more naturally occurring English, which is called “reduced forms” [6]. Reduced forms refers to basic elements of this naturally occurring spoken English, integral and pervasive elements of spoken English, that are seriously neglected in phoneme modeling.

In the language recognition task, the phone recognizer is commonly based on the same structure with that is used in speech recognition (SR) system. Actually, the requirements for the phone recognizers in PLR and SR are different due to the difference of the recognizing ranges and goals. Speech recognition is a language-dependent task, the relationship of phonemes is relatively less than that in language recognition. So the acoustic model is acquired to contain as much states as possible to describe the specific language more precisely, which limits the generalizing ability of the acoustic model. While the language recognition is a language-independent task, the phone recognizer needs to carry on in many languages of the phone recognizing task, which requests high consistency and the robustness of phone recognizing in the different language.

In this paper, an acoustic model is proposed specially for the language recognition tasks to solve the problems mentioned above, whose building process contains dictionary merging, phoneme splitting, phoneme clustering, state clustering and DNN-HMM acoustic modeling (DPPSD). DPPSD acoustic modeling method takes account on the reduced forms of the English spoken training materials and calculates the distance measure at the state-level, which is more accurate than other methods based on the phone-level alignment such as the acoustic likelihood method, can get a phone inventory that achieves more superior language recognizing performance.

The remainder of the paper is organized as follows: Section 2 shows the difference between the DPPSD English acoustic model and the traditional English acoustic model in formulation and

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

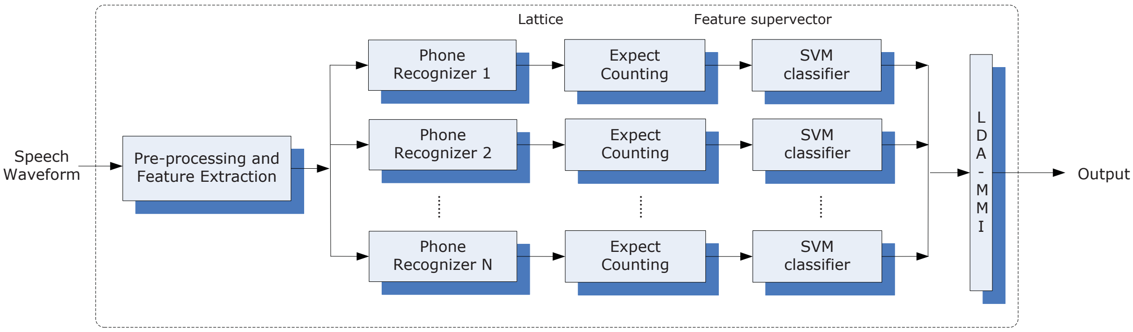

Figure 1. Architecture of PPR-VSM system.

the DPPSD phonotactic language recognition system. Section 3 shows the experimental setup. To evaluate the performance of the proposed DPPSD acoustic model, experiments are carried out and the results are analyzed in Section 4. Finally, Section 5 concludes the paper.

II. DPPSD Acoustic Model Based Phone Recognizer Specially for Phonotactic Language Recognition

DPPSD phone recognizer front-end based phonotactic language recognition system is shown in Figure 2. The building process contains English dictionary merging, phoneme splitting, phoneme clustering, state clustering and DNN-HMM acoustic modeling.

A. English Dictionary Merging

In traditional speech recognition, the CMU [7] dictionary or TIMIT dictionary [8] is often used to define the standard English pronunciation. The CMU dictionary contains a total of 133,354 words' pronunciations labelled by ARPAbet symbol set, which is developed by the Advanced Research Projects Agency (ARPA). A total of 39 phonemes are contained in ARPAbet symbol set, in which are 24 consonants and 15 vowels. The TIMIT dictionary contains a total of 6,229 words' pronunciations labelled by American English pronunciation dictionary symbols developed by Kenyon and Knott of the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, which contains 27 consonants and 18 vowels in 45 phonemes.

There are two sides of the problem need to take into account when developing the phoneme set for the English phone recognizer for language recognition. On one hand, the more accurate description of the English phonemes are involved in the training speech, the more accurate acoustic model can be built; On the other hand, the expensive cost of organizing training resources such as training data annotation and dictionaries must be considered, so existing standard pronunciation dictionaries and English voice data annotation resources are needed to make maximum use. The TIMIT phonetic description of the various pronunciations are more subtle than the CMU phonetic transcription but contains only 6,229 words' pronunciations. The rich resource of CMU dictionary indicates that it can be an effective compensation for the TIMIT dictionary. In this paper a refined dictionary is developed using the detailed features of CMU dictionary words and TIMIT phoneme sets to describe phonemes with more standard pronunciation definitions.

The improved pronunciation dictionary of English phone recognizer in this paper is based on the CMU dictionary. The difference between CMU dictionary and TIMIT dictionary is a small number of phonemes in the CMU dictionary that have different corresponding relationships with TIMIT phonemes according to the syllable structure and the change in stress, and the remaining phonemes correspond to TIMIT dictionary phonemes one by one. Therefore, the pronunciation dictionary of English phone recognizer in this paper is constructed by putting the same phonemes of TIMIT and CMU together and subdividing the phonemes according to reduced form of spoken English.

B. Reduced Form of Spoken English and Phone Splitting

All languages have this type of variation from written to spoken texts: "It results from a simple law of economy, whereby the organs of speech, instead of taking a new position for each sound, tend to draw sounds together with the purpose of saving time and energy" . With English, this process of assimilation is combined with contractions, elision, and reduction to produce the connected speech commonly referred to as "reduced forms". Naturally occurring English conversation, whether formal or informal, fast or slow, is full of these reduced forms. This creates a serious challenge for building acoustic model. In practical applications, it is necessary to adjust the pronunciation dictionary according to the reduced forms of spoken English, divide the phonemes that have the reduced pronunciation into another independent phonemes and build acoustic model for them to describe the syllable pronunciation more accurately.

C. Phoneme clustering

In this paper a State-Time-Alignment (STA) [9] phoneme clustering method is proposed to balance the performance and the complexity of the phone recognizer. There are two kinds of popular methods widely used, one is to search an universal and compact phone inventory by using phonetic knowledge [10] such as the International Phonetic Alphabet (IPA), the other is to merge the phone models using data-driven clustering method based on model-distance measures such as acoustic likelihood [11] or Bhattacharyya [12] distance. These methods have been used in Automatic Speech Recognition (ASR) systems and improve the performance of ASR, however, they are not suitable for language recognition. For language recognition work involves many type of languages, it is very important to determine an accurate phone inventory for the phone recognizer front-end of the language recognition system. STA method calculates the

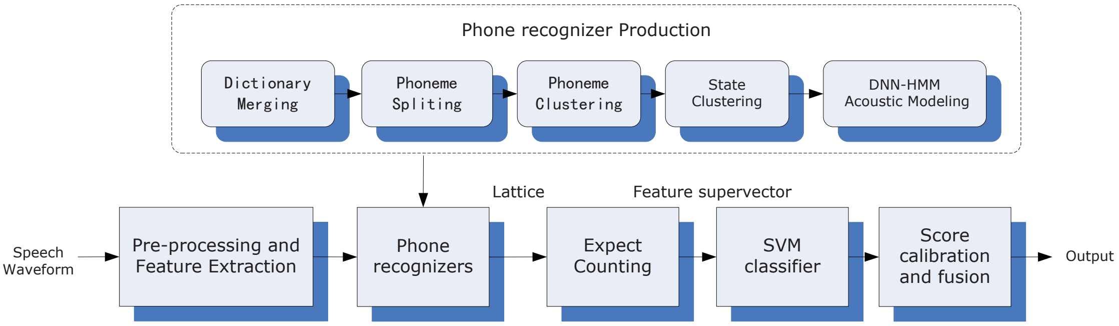

Figure 2. Flowchart of English phone recognizer production.

distance measure at the state-level, which is more accurate than other methods based on the phone-level alignment such as the acoustic likelihood method, that can get a phone inventory that achieves more superior language recognizing performance.

In STA clustering method, the distance between phone model p and q are defined as:

$$\begin{aligned} &D\left(p,q\right) \\ &= \tfrac{1}{2}\left(\sum_{j,k} c\left(p_j,q_k\right) D\left(p_j,q_k\right) + \sum_{j,k} c\left(q_j,p_k\right) D\left(q_j,p_k\right)\right), \end{aligned} \tag{1}$$

here $c\left(p_j,q_k\right)$ is the count of times t where model p has its state j active and model q has its state k active, which is calculated as:

$$c\left(p_j,q_k\right) = \sum_{i}\sum_{t=1}^{T_i} \gamma_{p_j}(t)\gamma_{q_k}(t), \tag{2}$$

in which i is is the sub-segment index, T_i is the ending time of each sub-segment, $\gamma_{p_j}(t)$ and $\gamma_{q_k}(t)$ are occupation probabilities (0 or 1) defined in the forward-backward algorithm.

$D(p_j,q_k)$ is the distance of the jth state of model p and the kth state of model q in sub-segment by the Bhattacharyya distance measure, which is calculated as:

$$\begin{aligned} &D\left(p_j,q_k\right) \\ &= \tfrac{1}{8}\left(\mu_{p_j}-\mu_{q_k}\right)^T\left[\tfrac{\Sigma_{p_j}+\Sigma_{q_k}}{2}\right]^{-1}\left(\mu_{p_j}-\mu_{q_k}\right) \\ &+\tfrac{1}{2}\ln\frac{\left|\frac{\Sigma_{p_j}+\Sigma_{q_k}}{2}\right|}{\sqrt{\left|\Sigma_{p_j}\right|\left|\Sigma_{q_k}\right|}}, \end{aligned} \tag{3}$$

where μ_{p_j} and $\sum_{p_j}$ denote mean vector and variance of the jth Gaussian state of phone model p.

In STA processing, we merge two phone models according to the minimum distance, in which the parameters of model l are updated as:

$$\mu_{l_y} = \frac{m_{p_y}\mu_{p_y} + m_{q_y}\mu_{q_y}}{m_{p_y}+m_{q_y}}, \tag{4}$$

$$\sum\nolimits_{l_y} = \frac{m_{p_y}({\mu_{p_y}}^2+\sum_{p_y})+m_{q_y}({\mu_{q_y}}^2+\sum_{q_y})}{m_{p_y}+m_{q_y}} - {\mu_{l_y}}^2 \tag{5}$$

here m_{p_y} and m_{q_y} denotes the data statistics of yth state of model p and model q. The merging process is repeated until desired number phone classes is reached. All the phone models are turning into new DPPSD phoneme set after phoneme clustering. The dictionary, questions for decision tree and transcription files for training data are all processed based on the clustering information.

D. DPPSD Phone Recognizer Front-end Based Phonotactic Language Recognition

After the DPPSD acoustic modeling, the input data x is mapped from input space $\mathcal{X}$ into a new high dimensional DPPSD feature space $\mathcal{F}_{\text{DPPSD}}$: $\Phi : \mathcal{X} \rightarrow \mathcal{F}_{\text{DPPSD}}$, and then linear machines is builded to classify in the feature space.

Then training or test utterance x is mapped into the DPPSD feature space as:

$$\mathbf{\Phi}_{\text{DPPSD}} : x \rightarrow \varphi_{\text{DPPSD}}(x). \tag{6}$$

and the N-gram feature supervectors $\varphi(x)$ is calculated as:

$$\varphi_{\text{DPPSD}}(x) = [p(d_1|\ell_x), p(d_2|\ell_x), ..., p(d_F|\ell_x)], \tag{7}$$

here d_q is the DPPSD phone inventory based N-gram phoneme string $d_q = s_q...s_{q+n-1}$ $(n = N)$ and $F = f^N$ (f denotes the size of the DPPSD based phone inventory). ℓ_x denotes the DPPSD phone inventory based lattice converted from data x. $p(d_q|\ell_x)$ is the probability of the N-gram $s_q...s_{q+N-1}$ in the lattice ℓ_x.

Also the probability of the phone sequence $s_q...s_{q+N-1}$ in the DPPSD phone inventory based lattice is calculated as follows:

$$p(s_q...s_{q+N-1}|\ell_x) = \frac{c(s_i...s_{i+N-1}|\ell_x)}{\sum_{\forall m} c(s_m...s_{m+N-1}|\ell_x)}, \tag{8}$$

Given the DNN-HMM based DPPSD acoustic model Λ_{DNN}, the expected counts over all possible hypotheses in the lattice of speech utterance x are computed as follows [13]:

$$\begin{aligned} &c(s_i, ..., s_{i+N-1}|\ell_x) \\ &= E[c(s_i, ..., s_{i+N-1})|X, \Lambda_{DNN}, M'] \\ &= \sum_{s_i...s_{i+N-1}\in\ell_x} [\alpha(s_i)\beta(s_{i+N-1})\prod_{j=i}^{i+N-1}\xi(s_j)], \end{aligned}$$

where M is the estimates of the N-gram probabilities that maximize $\sum_H f(X|H, \Lambda_{DNN})P(H|\mathcal{L})$ $(H = s_i...s_{i+N-1}$, $\mathcal{L}$ is the language under consideration, $f(X|H, \Lambda_{DNN})$ is the likelihood of the speech utterance X given $\mathcal{L}$ and H). $\alpha(s_i)$ is the forward probability of the starting node of $s_i...s_{i+N-1}$ and $\beta(s_{i+N-1})$ is the backward probability of the ending node of $s_i...s_{i+N-1}$. $\xi(s_j)$ denotes the posterior probability of the edge s_j.

Then the SVM output score is calculated as follows:

$$\begin{aligned} &f(\varphi_{\text{DPPSD}}(x)) \\ &= \sum\nolimits_v \alpha_v K_{\text{TFLLR}}(\varphi_{\text{DPPSD}}(x), \varphi_{\text{DPPSD}}(x_v)) + d, \end{aligned} \tag{9}$$

here $\varphi_{\text{DPPSD}}(x_v)$ are support vectors. The TFLLR kernel is calculated as [14]:

$$\begin{aligned} & K_{\text{TFLLR}}(\varphi_{\text{DPPSD}}(x_i), \varphi_{\text{DPPSD}}(x_j)) \\ & = \sum_{q=1}^{F} \frac{p(d_q|\ell_{x_i})}{\sqrt{p(d_q|\ell_{all})}} * \frac{p(d_q|\ell_{x_j})}{\sqrt{p(d_q|\ell_{all})}}, \end{aligned} \tag{10}$$

Then the posterior probabilities of all the belief score vector are maximized using the LDA-MMI algorithm [15] with objective function as follows [16]:

$$F_{\text{MMI}}(\lambda) = \sum_{\forall i} \log \frac{p(\boldsymbol{x}_i|\lambda_{g(i)})P(g(i))}{\sum_{\forall j} p(\boldsymbol{x}_i|\lambda_j)P(j)}, \tag{11}$$

here $\boldsymbol{x} = [f(\varphi_{\text{DPPSD}}(x))]$ and $g(i)$ indicates its class label. $P(j)$ is the prior probability of class j. $p(\boldsymbol{x}|\lambda)$ is weighted Gaussian mixtures that describe a general distribution:

$$p(\boldsymbol{x}|\lambda) = \sum\nolimits_{\forall m} \omega_m \mathcal{N}(x; \mu_m, \Sigma_m), \tag{12}$$

here $\mathcal{N}(\cdot)$ denotes the normal distribution with a parameter set that is often referred as $\lambda = \{\omega_m, \mu_m, \Sigma_m\}$. Here μ_m, Σ_m and ω_m are the mean vector, covariance matrix and the weight of the m-th Gaussian mixture.

Such acoustic model has three advantages. First, STA method calculates the distance measure at the state-level, which is more accurate than other methods based on the phone-level alignment such as the acoustic likelihood method, that can get a phone inventory that achieves more superior language recognizing performance. Second, DPPSD acoustic model adds dictionary merging, phoneme splitting, phoneme clustering, state clustering in building processing than the traditional acoustic model, while it costs no more computation in decoding but gets a remarkable improvement. Third, DPPSD based phone recognizer can explore more discriminative information than traditional phone recognizer, then more effective information can be extracted for the language recognition system to classify.

III. Experimental setup

A. System setup

A PR-SVM language recognition system is used as baseline system in this paper. The first step is to tokenize speech by the means of running phone-recognizers and the decoder named HVite produced by HTK [17] is used to produce phone lattices, and an open software named lattice-tool (SRILM) [18] is used to provide the posterior probabilities of the phone occurrences to produce phone counts. Then, a popular classifier LIBLINEAR [19] is employed to classify the feature supervector. Finally, LDA-MMI algorithm [20] is used for score calibration among the different acoustic models.

The DPPSD English DNN-HMM acoustic model is trained using about 100 hours' Switchboard English corpus [21]. The same training algorithm are applied to train DNNs as in [22] in this work. Input features to DNNs in the training stage are 13-dimensional PLP features plus their first and second order derivatives, which are all normalized to have zero mean and unit variance based on conversation-side information [23]. A CUDAMat library [24] is used as the implementations of the DNN.

B. Training, test and developing database

About 180,000 training data used in this paper belongs to: (1) the Call-Home Corpus; (2) the OHSU Corpus provided by NIST LRE 2005; (3) the VOA Corpus; (4) the OGI Corpus; and (5) the Call-Friend Corpus.

The experiments are conducted on the test trials of the National Institute of Standards and Technology Language Recognition Evaluation 2009 (NIST LRE 2009) tasks. The test database comprises 41793 test segments of 23 languages for 30-s, 10-s, and 3-s nominal duration test.

22701 conversations are selected from the database provided by NIST for the 2003, 2005 and 2007 LRE and VOA for developing purposes.

C. Evaluation measures

The performance of language recognition system in this paper is reported by Equal Error Rate (EER) and average cost performance (C_{avg}), which are defined by NIST LRE 2009 [25].

IV. Experimental results and analysis

In this section, we present our experimental results of the language recognition system using English phone recognizer on the NIST LRE 2009. More than 1,800,000 utterances are used for training.

A. Comparison of different kinds of dictionaries

The performance of different kinds of dictionary based language recognition system are investigated in this subsection and the EER and C_{avg} of the different front-end language recognition system are listed in Table I, in which the dictionary of 39 phonemes is CMU and the dictionary of 45 phonemes is TIMIT.

It can be seen from the results of Table I that the number of the phonemes of the dictionary is varied from 39 to 58. For each fixed frontend, the performance of the language recognition system first decreases and then increases with the decreasing of number of the phonemes inventory of the acoustic model, and the best performance occurs at the phone inventories of size 47. In fact the performance of the 47 phonemes' DPPSD English dictionary based phonotactic language recognition outperforms that of both CMU and TIMIT based system, which means the balance point of generalization and accuration of the acoustic model. Notice that the performance of the 47 phonemes' DPPSD English dictionary based phonotactic language recognition also outperforms the systems whose acoustic model using knowledge-based and data-driven phoneme clustering method. So when 47 phonemes DPPSD set English dictionary is adopted and the language recognition system achieves the best performance.

B. Effects of the number of GMMs' states

Table II shows the language recognition performance of different number of GMMs' states based English DNN-HMM acoustic model. The English GMM-HMM triphone acoustic model contains from 144 to 9308 [23] states with 32 Gaussians each. Note that 9308 GMMs' states acoustic model corresponds to the acoustic model used in speech recognition system. Better performance is achieved using 154-states DNN-HMM acoustic model than that of 258-states, 904-states and 9032-states, that means the 154-states acoustic model is more generalizing and robust in recognizing phonemes of different languages than

Table I
PERFORMANCE OF DIFFERENT KINDS OF DICTIONARY BASED LANGUAGE RECOGNITION SYSTEM, NIST LRE 2009, (EER/C_{AVG} IN %).

Number of phonemes		30s	10s	3s
		EER	EER	EER
Baseline System	39(CMU)	2.37	7.04	21.42
	45(TIMIT)	2.23	6.84	20.20
Phoneme Clustering	knowledge-based	2.19	6.99	20.26
	data-driven	2.13	6.80	20.17
STA	58	2.29	6.93	20.62
	53	2.26	6.81	20.12
	51	2.18	6.69	20.15
	47	**2.09**	**6.66**	**19.72**
	46	2.13	6.68	20.04

other acoustic models, even than the acoustic model used in speech recognition system. Therefore, English DNN-HMM acoustic model with 47 phoneme DPPSD dictionary and sigmoidal networks of 154 GMM-HMM states are used for all the following experiments.

Table II
PERFORMANCE OF DNN-HMM ACOUSTIC MODEL WITH DIFFERENT GMMS' STATES, NIST LRE 2009 (EER/CAVG IN %).

Number of GMMs' States	30s	10s	3s
	EER	EER	EER
9032	2.70	8.45	23.83
904	2.42	7.33	21.24
258	2.27	6.97	20.11
154	**2.09**	**6.66**	**19.72**

C. Comparison of different kinds of acoustic model

Table III compares the results of language recognition system using an ANN-HMM model, a GMM-HMM model with 154 states and a DNN-HMM model, which are all trained using the 100-hour subset of English Switchboard corpus. ANN-HMM acoustic model is trained using TRAP feature and a context window of 21 frames. Experiments show that DNN-HMM acoustic model provide dramatic improvements in language recognition accuracy and offers a relative EER reduction of 28.42%, 14.06%, 18.70% over the ANN-HMM acoustic model, and a relative EER reduction of 12.55%, 7.20%, 2.47% over the GMM-HMM acoustic model for 30s,10s and 3s, respectively. The performance of longer speech utterances (30s) improves more dramatically than that of 10s and 3s because DNNs are more powerful in modeling long context acoustic events than GMM-HMMs. Figure 3 shows the DET curves for NIST LRE 2009.

D. Comparison of real time factor for decoding

Table IV shows decoding real time factor of TIMIT, CMU and DPPSD acoustic model. Although the training cost of DPPSD acoustic model is a little expensive compared with training TIMIT and CMU acoustic model, decoding in DPPSD acoustic model is very efficient because the structure of DNN-HMM acoustic model is same with TIMIT and CMU acoustic model, the decoding real time factor only increases with the increasing of the number of phoneme inventory.

Table III
PERFORMANCE OF DIFFERENT KINDS OF ACOUSTIC MODEL BASED SYSTEMS, NIST LRE 2009, (EER/C_{AVG} IN %).

Acoustic model of phone recognizer	30s	10s	3s
	EER	EER	EER
EN-GMM-HMM (39 phonemes)	2.54	7.88	21.20
EN-GMM-HMM (47 phonemes)	2.39	7.20	20.22
EN-ANN-HMM (39 phonemes)	3.13	8.30	25.52
EN-ANN-HMM (47 phonemes)	2.92	7.75	23.41
EN-DNN-HMM (39 phonemes)	2.37	7.04	21.42
EN-DNN-HMM (47 phonemes)	**2.09**	**6.66**	**19.72**
EN fusion	1.39	4.28	15.83

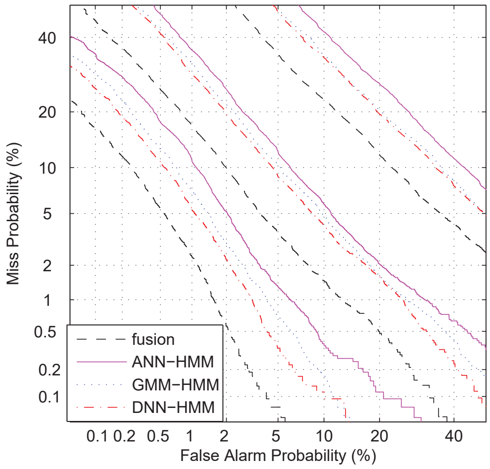

Figure 3. DET curves for NIST LRE 2009, English phone recognizer frontend.

Table IV
COMPARISON OF REAL TIME FACTOR FOR DECODING, NIST LRE 2009, 30-S TEST. CPU: XEON E5520@2.27 GHZ, RAM: 8 GB, SINGLE THREAD. GPU: GEFORCE GTX 275, RAM: 1 GB, 240 CUDA CORE.

acoustic model	TIMIT	CMU	DPPSD
RT factor	0.069	0.064	0.071

V. CONCLUSION

This paper has presented how to build a dramatic English acoustic model specially for phone recognizer front-end of phonotactic language recognition. The DPPSD English acoustic model is generalizing and robust to different languages, which is more suitable for language recognition than that is used in speech recognition and increases the performance of the system on accuracy without sacrificing its structure simplicity and computational effort. The experimental results evaluated on NIST LRE 2009 tasks have confirmed that the proposed DPPSD English acoustic model based system yields an EER of 2.07%, 6.66% and 19.22%, which achieves a 11.81%, 5.39% and 7.94%

relative deduction for 30s, 10s and 3s, respectively compared with the traditional CMU dictionary based English acoustic model, and a 6.28%, 2.63% and 7.94% relative deduction for 30s, 10s and 3s, respectively compared with the traditional TIMIT dictionary based English acoustic model.

As for future work, we will develop effective adaptation techniques of DNN and use DPPSD method to build multilingual phone recognizer. More work needs to be done, especially, in the direction of understanding which TDNN architectures would work best for phone recognition, and how to find such architecture, including determining good loss and learning rate, activation functions, cost optimizers, etc.

ACKNOWLEDGMENTS

Project supported by the National Natural Science Foundation of China (No.61370034, No.61403224 and No.61273268). Ya-Nan Li is the corresponding author.

REFERENCES

[1] Zissman, M.A.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing **4**(1), 33–44 (1996)

[2] Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller, J.R.: Approaches to language identification using gaussian mixture models and shifted delta cepstral features. Proc. ICSLP, 33–36 (2002)

[3] Li, H., Ma, B., Lee, K.-A.: Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE **101**(5), 1136–1159 (2013)

[4] Sim, K.C., Li, H.: On acoustic diversification front-end for spoken language identification. IEEE Trans. on Audio, Speech and Language Processing **16**(5), 1029–1037 (2008)

[5] Schwarz, P.: Phoneme recognition based on long temporal context (2009)

[6] Brown, J.D., Hilferty, A., Enright, S.: Listening for reduced forms. Tesol Quarterly **20**(4), 759–763 (2012)

[7] University, C.M.: Cmu pronouncing dictionary

[8] Zue, V.W., Seneff, S.: Transcription and alignment of the timit database (1996)

[9] Qian, Y., Jia, L.: Phone modeling and combining discriminative training for mandarinenglish bilingual speech recognition. In: IEEE International Conference on Acoustics Speech & Signal Processing (2010)

[10] Yu, S., Zhang, S., Bo, X.: Chinese-english bilingual phone modeling for cross-language speech recognition. In: IEEE International Conference on Acoustics (2004)

[11] Hler, J.: Multilingual Phone Models for Vocabulary-independent Speech Recognition Tasks, (2001)

[12] Mak, B., Barnard, E.: Phone Clustering Using the Bhattacharyya Distance, (1996)

[13] Gauvain, J.L., Messaoudi, A., Schwenk, H.: Language Recognition Using Phone Lattices. In: Proc. ICSLP, Jeju Island, pp. 1283–1286 (2004)

[14] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: Phonetic speaker recognition with support vector machines. Advances in neural information processing systems **16** (2003)

[15] Matejka, P., Burget, L., Glembek, O., Schwarz, P., Hubeika, V., Fapso, M., Mikolov, T., Plchot, O.: BUT system description for NIST LRE 2007. In: Proc. 2007 NIST Language Recognition Evaluation Workshop, pp. 1–5 (2007)

[16] Povey, D.: Discriminative training for large vocabulary speech recognition. Cambridge, UK: Cambridge University **79** (2004)

[17] Young, S., et al: The HTK book. Cambridge University Engineering Department **3** (2002)

[18] Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. [Online].Available: http://www.speech.sri.com/projects/srilm/ (2002)

[19] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research **9**, 1871–1874 (2008)

[20] Zhang, W.-Q., Hou, T., Liu, J.: Discriminative score fusion for language identification. Chinese Journal of Electronics **19**, 124–128 (2010)

[21] Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference On, vol. 1, pp. 517–520 (1992). IEEE

[22] Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on **20**(1), 30–42 (2012)

[23] Cai, M., Shi, Y., Liu, J.: Deep maxout neural networks for speech recognition. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop On, pp. 291–296 (2013). IEEE

[24] Mnih, V.: Cudamat: a cuda-based matrix class for python. Department of Computer Science, University of Toronto, Tech. Rep. UTML TR **4** (2009)

[25] NIST: The 2009 NIST language recognition evaluation plan. In: Http://www.itl.nist.gov/iad/mig/tests/lang/2009/ (2009)