High-Resolution Acoustic Modeling and Compact Language Modeling of Language-Universal Speech Attributes for Spoken Language Identification

Yannan Wang[1], *Jun Du*[1], *Lirong Dai*[1], *Chin-Hui Lee*[2]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, P. R. China
[2] School of Electrical and Computer Engineering, Georgia Institute of Technology, USA

wyn314@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a framework to automatically construct a collection of high-resolution (HR) language-universal units for spoken language identification (LID). Based on the popular phone recognition language modeling (PRLM) approach to LID, a set of universal attribute recognizers (UARs) is first established to replace phone recognizers (PRs) using manner and place of articulation as attribute units and context-dependent (CD) attribute models are then built to achieve high-performance attribute transcription. To alleviate the difficulty of data sparsity in n-gram language modeling (LM) of these CD units, a clustering algorithm is proposed to compact the number of utilized attribute units in LM. Tested on the 2009 National Institute of Standards and Technology Language Recognition Evaluation for the 30-sec task using the same English Switchboard-I training data for acoustic modeling, our proposed approach achieves an equal error rate (EER) of 2.34%, representing a relative EER reduction of over 20% from the results of 2.88% obtained with the conventional PRLM techniques. To the best of our knowledge, this is the first time a single UAR based LID system significantly outperforms a signal PR based system with the same set of training data from a single language.

Index Terms: spoken language identification, automatic speech attribute transcription, manner and place of articulation, high-resolution modeling, phone recognition language modeling

1. Introduction

Spoken language identification (LID) is a process of determining the language identity of a speech segment spoken by an unknown speaker. For text-independent language identification task, there are two main categories of information used in the recognition procedure of humans: the prelexical information and the lexical semantic knowledge [1], such as acoustic phonetics, prosody, phonotactic structure, vocabulary and so on. Among all the information sources, the acoustic and phonotactic features are the optimal choices. The acoustic feature mainly refers to the physical sound pattern [2] difference while the phonotactic feature is represented by the combination laws of phones in different languages. Based on the two types of features there are two main approaches to LID: the acoustic and phonotactics approaches [2, 3, 4, 5]. The former approach employs only acoustic features to generate feature vectors which are used to train models, such as Gaussian mixture models (GMMs) [6]. The latter approach makes use of linguistic information and acoustic phonetics information. One of the widely used phonotactics approaches is the phone recognition followed by language modeling (PRLM) [5] which consists of the tokenizer front-end to obtain token sequences and the n-gram language model back-end to deliver the LID decision. Recently vector space modeling approach has also been proven effective to produce competitive LID results [7, 8, 9].

In this paper, we focus on another phonotactics approach based on speech units representing articulation features to build a set of language-independent tokenizers collectively called universal attribute recognizers (UARs), which have been utilized in a recently proposed automatic speech attribute transcription (ASAT) paradigm [10, 11] for automatic speech recognition (ASR). The features, including manner of articulation and place of articulation, are universally adopted across languages [12] and therefore there are two advantages. First, it alleviates the problem of missing phones in the front-end phone recognizer of the PRLM systems and enhances the capability of models by sharing data from different languages [5, 13]. Second, the size of the attribute inventory is usually smaller than that of various phone inventories. Moreover we can improve the transcription accuracy with high-resolution (HR) models beyond context-independent (CI) attribute models to deliver better LID performance [10, 8, 9].

The right-context (RC) dependent acoustic models [14, 15] have been used on the NIST 2003 spoken language evaluation task [16]. In this study, we propose high-resolution (HR) models which exploit both left and right attribute context information [17]. Without any constraint, the number of units can be more than one thousand. Our experiments show that too many HR units can lead to a degradation of LID performance, which can be explained in two aspects. First, the training data will not be sufficient to produce the same high-accuracy model for all units in the collection. Second, unlike the RC models in which the token sequence after decoding still consists of CI units, our HR units are originally in a CI form which can result in an infeasible n-gram language model if the number of HR units is large. To address these issues, we put forward a clustering algorithm mainly based on the distribution of the training data to generate a compact set of HR attribute units. Tested on the 2009 National Institute of Standards and Technology Language Recognition Evaluation for the 30-sec task with the same English Switchboard-I training data for acoustic modeling, our proposed UAR approach achieves an equal error rate (EER) of 2.34%, representing a relative EER reduction of over 20% from the LID results of 2.88% obtained with the conventional PRLM techniques. By fusing different UARs generated with both the English and Mandarin training data, we finally achieve EERs of 1.93%,3.6%,11.94% for 30s, 10s, and 3s test utterances, respectively. These EERs are reduced to 1.51%, 2.86% and 10.18% when Hungarian PR and Russian PR are also fused together.

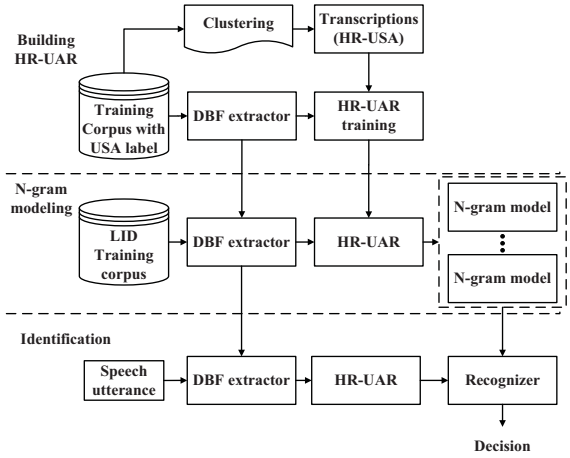

Figure 1: *A block diagram of the language identification system with HR language-universal speech attribute units.*

2. Speech Attribute Based LID Systems

2.1. LID System Overview

The architecture of our system is shown in Figure 1. For building HR-UAR front-end, firstly we employ the intensively used deep neural network (DNN) [18, 19] to train a feature extractor, namely deep bottleneck feature (DBF) [20, 21, 22] extractor. Then we extract the DBF of transcribed corpus to train our tokenizer front-end. Next, for the n-gram modeling stage, we also extract the DBF of training material of LID corpus in the same way. With the obtained DBF we make use of UAR to create the lattice of HR attribute units and employ the integrated language models [23] to conduct the training of the back-end in our system. Finally, with the n-gram models to approximate the probability distribution of the co-occurrences of the chosen HR attribute units we can compute the posterior probabilities that the test utterance belongs to a certain language and make the LID decision.

2.2. Modeling of Language-Universal Attribute Units

The set of language-universal speech attributes used in this study is the same as in [13], consisting of place and manner of articulation, some of them often referred to as distinctive features [24] commonly adopted to characterize acoustic phonetics [25] of speech sounds for all spoken languages [13]. With these universally defined units we can construct universal manner recognizers (UMRs) and universal place recognizers (UPRs) to convert phoneme-based into attributes-based transcriptions according to the mapping tables between the phonemes and attributes. Apart from these regular attributes, we also employ the ”silence” token to represent the soundless segments and ”noise” token to represent the noisy background fragments, respectively. Nonetheless the decoded noise tokens are ignored when building language models.

2.3. High-Resolution Speech Attributes

In phonotactics approach to LID, the phonotactic information which captures the language characteristics is usually conveyed by n-gram statistics. In our study we obtain up to 4-gram statistics from our decoding lattice [26] by a token recognizer. Intuitively, the number of tokens employed in the n-gram statistics vector is critical to the LID performance. It also determines the vector in n-gram modeling which may be beyond our processing capacity if the number of tokens is too large.

It has been demonstrated that the accuracy of tokenizer is one of the key factor in determining the LID performance [27]. With a broad class of phones, the higher accuracy obtained could compensate for the loss of detailed phone information [28]. On the Eval03 NIST detection test the RC UAR could achieve a competitive performance to conventional phone recognizer [14]. However, for the more difficult LRE09 task with much more target languages and easily confusable language pairs, the ability of UAR with manner-only or place-only configuration is inadequate. Accordingly, we present a new approach to build the HR attribute models of tokenizers. This approach takes both left and right context information across words. For instance, for an utterance which is represented by

$$t_1\ t_2\ t_3\ t_4\ ...\ t_{n-2}\ t_{n-1}\ t_n$$

where the t_i denotes the i-th attribute, the obtained HR attributes sequence is as follows:

$$t_1 - t_2 + t_3 \quad t_2 - t_3 + t_4 \quad ... \quad t_{n-2} - t_{n-1} + t_n$$

The new units introduced above are regarded as initial HR attributes. While capturing the context variabilities the above conversion procedure increases the size of attribute collection to more than one thousand. By contrast, the popular approach to implement context-dependent (CD) models using both left and right contexts is also explored in our work, which is an effective solution to elevate the resolution of acoustic models using context information. The main difference between our HR models and CD models lies in that the traditional CD UAR still focuses on the central units and uses them to dig into the phonotactic nature of languages while in our system we treat the HR units as some kind of new attributes. However, due to the large number of initial HR units, the data insufficiency becomes a non-trivial issue. Moreover we are not capable of constructing the high-order n-gram statistics with so many tokens. To address those problems we exploit a clustering strategy to control the number of HR attribute units.

2.4. Clustering of CD Attributes for Compact LM

With the initial HR attribute collection described in section 2.3, we adopt a data-driven method to obtain the final HR attribute units. This procedure is shown in Figure 2. First the CI attribute t_i in transcriptions is converted to HR attributes and then grouped according to the central attribute. In each group we sort them in ascending order of training transcription amount and after that we start the clustering procedure. In each iteration we have two steps:

(1) Search for the HR attribute unit $l_{i1} - t_i + r_{i1}$ with the minimum transcription amount of the training corpus.

(2) Merge the training transcription of the HR attribute unit $l_{i1} - t_i + r_{i1}$ with its neighbor $l_{i2} - t_i + r_{i2}$, and define a new attribute unit to replace these two attributes. Finally reorder this group of new HR attribute units.

With each iteration the number of the HR attribute units decreases by one. By repeating the above procedure we can get a collection of attribute units with a predefined size K. In this approach the size of final attribute units inventory is crucial to LID

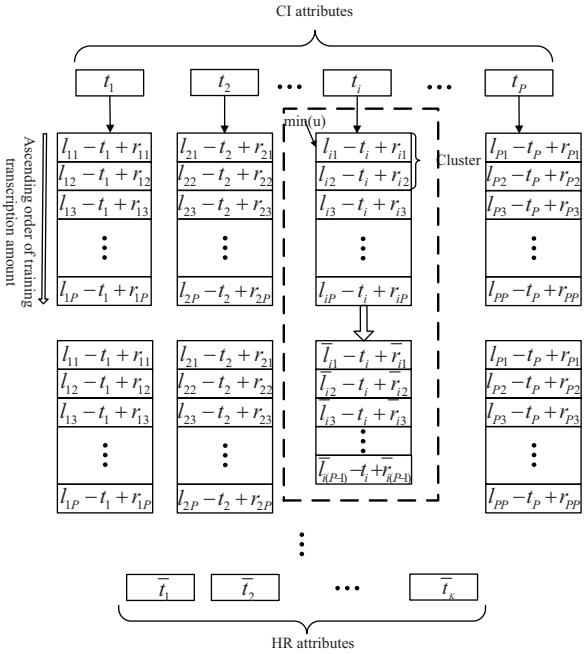

Figure 2: *Clustering procedure to generate the HR attributes.*

performance. To design an automatic approach to determine K, we compute a ratio at each iteration:

$$R = \frac{min(u)}{N/P} \tag{1}$$

where N is the number of all attribute transcriptions in the training corpus, P is the number of attribute inventory defined in [13], $min(u)$ is the number of the least attribute transcription in the current iteration. We stop the clustering procedure when R reaches to a predefined threshold (5% in this study). Our clustering algorithm is very simple and only the training transcriptions are used, which is quite different from the traditional unit merging strategy in which acoustic models of all units should be trained to measure the similarity among the units according to the acoustic distances. In this way our method also benefits from the balance of training data of different units which is important in consideration of the larger size of our inventory.

3. LID Experiments and Result Analysis

3.1. Experimental Setup

In this work 309-hour Switchboard-I training set and the in-house corpus of 1000-hour Mandarin telephone conversational speech were used to train the articulatory recognizer, respectively. For the training of DBF extractor the input of the deep neural network (DNN) was the 43-dimensional feature vector consisting of 13-dimensional perceptual linear prediction (PLP) feature [29] plus their first and second order derivatives and 4-dimensional feature vector related to the pitch and the confidence coefficient of voiced or unvoiced of current frame. Moreover, the temporal window size of input PLP feature was set to 21 with 10 frames for both left and right windows. This DNN had 5 hidden layers and there were 2048 units for each hidden layer except for the bottleneck layer with 55 units. We implemented the UARs within the HMM framework [30] trained with maximum likelihood (ML) criterion [31] and the HMM state probability density function was estimated with Gaussian mixture model (GMM). For each of the HR attribute units we employed a 5-state HMM to model it, and the Gaussian mixture number of each state was 80.

Our LID experiments were conducted on the LRE09 dataset which included 23 target languages [32]. The utterances in the training set were recorded with two channels, namely Conversational Telephone Speech (CTS) and narrowband Voice of America (VOA) [32]. In our experiments we only made use of a subset of 15 hours speech data for each target language due to the imbalance of training data. And the selected data were split into the training set and development set. There were about 80 segments of 30s duration for each language in the development set. The test set consisted of three tasks according to different duration time, namely 3s, 10s and 30s.

3.2. Comparison between Different UARs

In this subsection we investigated the performance of different tokenizers which were designed as described above and the results were depicted in Table 1. In this table the phone recognizers refer to the Temporal Patterns Neural Network (TRAPs/NN) [33] recognizer for Hungarian (HU) and Russian (RU) developed by the Brno University of Technology (BUT) [34] who have been widely adopted as baseline system in LID tasks and the one trained on the bottleneck features of the Switchboard (EN-PR) corpus. It has been demonstrated that better LID performance can be delivered by improving the UAR acoustic resolution in [14] which were also proven in this table. Compared with straightforward CI-based systems, the RC dependent attribute models improved the LID performance indeed, e.g., EER declining from 12.75% to 8.32% using UPR for 30sec test utterance, while they still behaved very poorly. Similarly, for 30s task, the CD-UPR (at EER of 6.88) showed great advantages compared with the RC-UPR (at EER of 8.32) but it was still not as competitive to our proposed HR-UPR (at EER of 2.34). As for UMR, similar trend of promotions could be observed that the EER jumped from 6.97% to 2.54% for 30s evaluation task when switching from RC-UMR system to HR-UMR system.

Moreover, in Table 1 we found that the HR-UAR delivered lower EERs than those for our proposed EN-PR with about 20% average relatively reduction even though they were trained with the same speech corpus. With respect to the conventional HU-PR, the improvements were also remarkable from 2.62% dropping to 2.34% and 2.54% for HR-UPR and HR-UMR, respectively. Figure 3 plots the DET curves [35] of the LID results of HR-UAR and HU-PR on the LRE09 tasks. We could observe a clear wide gap between our HR-UAR based system and the HU PR system, especially for short time task, indicating the potential of DBF on short time tasks.

3.3. Size Selection of CD Attribute Collection

As mentioned above, the clustering procedure to reduce the number of HR attribute units is crucial to improve the LID performance. For the 309-hour Switchboard-I training set, the final sizes of attribute unit collection for UPR and UMR determined by Equation 1 were 84 and 86, respectively. In addition to the generated attribute collection as described above, we also adopted some other sizes of HR attributes as comparisons, such as $K = 40, 80, 120$. Their performances were shown in Table 2. From this table we can observe that the size of the HR attribute unit collection had a great influence on the LID result. Clearly, for both UMR and UPR, EERs of all duration

Table 1: *Performance comparison of different EN-PR, EN-UMR and EN-UPR tokenizers on LRE09 (EER and C_{avg} in %).*

		30s		10s		3s	
		EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
HU-PR		2.62	2.62	6.65	6.62	18.88	18.82
RU-PR		2.42	2.4	6.42	6.38	18.92	18.70
EN-PR		2.88	2.87	6.62	6.60	19.6	19.43
CI	UPR	12.75	12.53	20.35	20.27	32.71	32.56
	UMR	9.84	9.78	19.56	19.31	33.80	33.64
RC	UPR	8.32	8.30	18.56	18.52	31.12	31.15
	UMR	8.21	8.20	18.31	18.29	31.03	31.01
CD	UPR	6.88	6.85	15.56	15.55	29.70	29.48
	UMR	6.97	6.96	16.39	16.30	30.95	30.89
HR	UPR	2.34	2.34	5.32	5.30	16.13	16.05
	UMR	2.54	2.5	5.31	5.29	16.07	15.87

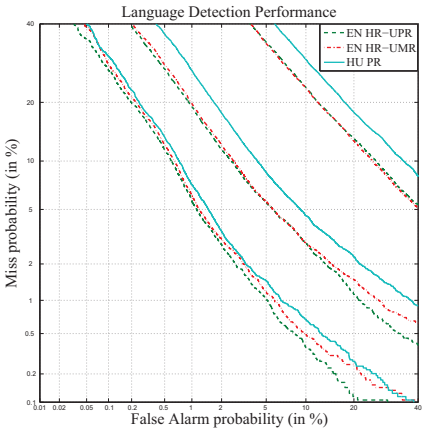

Figure 3: *DET curves for EN-UAR and HU-PR on LRE09.*

lengths for the test utterances reached to the minimum at 80 under the three manually determined configurations which verifies the degradation of LID performance with too many HR attribute units as mentioned above. With the size K automatically decided, UPR and UMR both performed competitively to their best configuration, e.g., for 30s task EER changing from 2.33% to 2.34% for UPR and from 2.63% to 2.54% for UMR. The complexity of our system is heavily affected by the attribute collection size during building language models. With K set to below 100, the efficiency problem is not severe.

3.4. Mandarin UARs and System Fusion

As shown in Table 1 the LID results of EN-UARs on the English training corpus with automatically decided CD unit set sizes was satisfying and achieved a great improvement over phone recognizers (EN-PR). Similarly, for the 1000-hour Mandarin training set, the determined sizes of the HR attribute collections for UPR and UMR were 112 and 97, respectively. The corresponding EER performances were listed in Table 3. The LID results were not so well as those on the English corpus, but the advantages of Mandarin UAR (MA-UAR) over Mandarin phone recognizer (MA-PR) are similar to those reported in Table 1 for English tokenizers.

Table 2: *EER(%) for different clustering size on LRE09 task using HR UPR and UMR on English corpus.*

		30s	10s	3s
EN HR-UPR	K=40	3.40	8.34	21.54
	K=80	2.33	5.29	16.03
	K=120	2.98	6.72	19.3
	Auto	2.34	5.32	16.13
EN HR-UMR	K=40	3.61	8.97	22.78
	K=80	2.63	5.91	16.87
	K=120	2.71	6.11	17.06
	Auto	2.54	5.31	16.07

As we know, only one phone recognizer can be constructed on one language in general. Resulting from the complementary nature of tokenizers on different languages, the parallel PRLM (PPRLM) system [5] is widely used. Moreover in our work we could combine two UAR tokenizers, i.e., UPR and UMR on a single-language corpus through a Gaussian back-end [36, 37]. Furthermore, when our proposed four UARs on English (EN) and Mandarin (MA) corpora were all fused, as shown in the next-to-bottom row in Table 3, the best ERRs of 1.93%, 3.6%, 11.94% were achieved. Finally, we fused the system based on HU-PR and our proposed HR-UARs (EN-UARs and MA-UARs) and obtained the EER results of 1.88%, 3.25%, 10.95% for 30s, 10s, and 3s test utterances, respectively, and EERs further fell to 1.51%, 2.86%, 10.18% as shown in the bottom row of Table 3 when the RU-PR was fused. These improvements, not surprisingly, owed in part to the complementary nature between fundamental speech attributes and phones [13] besides the various acoustic model architectures and features.

Table 3: *Fusion results for UARs on LRE09 task (EER in %).*

	30s	10s	3s
MA-PR	3.08	7.79	21.93
MA HR-UPR	2.63	7.02	21.25
MA HR-UMR	2.85	6.83	19.87
Fusion 4-UARs: EN-UARs + MA-UARs	1.93	3.6	11.94
Fusion: HU-PR + 4-UARs	1.88	3.25	10.95
Fusion: HU-PR + RU-PR + 4-UARs	1.51	2.86	10.18

4. Summary

We propose an attribute based LID approach with high-resolution acoustic models for token recognizer and compact language models. Tested on the LRE09 task we obtain a significant reduction in EER than conventional PRLM systems using the same training and testing data. For 30sec testing on the LRE09 task, we achieved an EER of 2.54% using a single EN-UAR. Furthermore, By fusing English UARs, Mandarin UARs, Hungarian PR, and Russioan PR, we achieved EERs of 1.51%, 2.86%, and 10.18% for 30s, 10s and 3s utterances, respectively.

5. Acknowledgment

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273264 and No. 61305002), the Electronic Information Industry Development Fund of China (Grant No. 2013-472), and the Programs for Science and Technology Development of Anhui Province, China (Grants No. 13Z02008-4 and No. 13Z02008-5).

6. References

[1] J. Zhao, H. Shu, L. Zhang, X. Wang, Q. Gong, and P. Li, "Cortical competition during language discrimination," *NeuroImage*, vol. 43, no. 3, pp. 624–633, 2008.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[3] A. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *Signal Processing Applications for Public Security and Forensics, 2007. SAFE '07. IEEE Workshop on*, April 2007, pp. 1–7.

[4] NIST language recognition evaluations. [Online]. Available: http://nist.gov/itl/iad/mig/lre.cfm

[5] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, p. 31, 1996.

[6] P. A. Torres-carrasquillo, E. Singer, M. A. Kohler, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002, pp. 89–92.

[7] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. on Audio, Speech, Language Process*, vol. 15, no. 1, pp. 271–284, 2007.

[8] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition." in *Proc. Interspeech*, 2009, pp. 168–171.

[9] ——, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[10] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[11] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition," in *Proc. Interspeech*, 2004, pp. 109–112.

[12] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. ICASSP*, 2008, pp. 4261–4264.

[13] Y. Wang, J. Du, L. Dai, and C.-H. Lee, "A fusion approach to spoken language identification based on combining multiple phone recognizers and speech attribute detectors," in *Proc. ISCSLP*, 2014, pp. 158–162.

[14] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition," in *Proc. Interspeech*, 2010, pp. 2718–2721.

[15] C.-H. Lee and B.-H. Juang, "A survey on automatic speech recognition with an illustrative example on continuous speech recognition of mandarin," *Computational Linguistics and Chinese Language Processing*, vol. 1, no. 1, pp. 01–36, 1996.

[16] A. Martin and M. Przybocki, "NIST 2003 language recognition evaluation." in *Proc. Eurospeech*, 2003, pp. 1341–1344.

[17] C.-H. Lee, L.-R. Rabiner, R. Pieraccini, and J.-G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, vol. 4, no. 2, pp. 127–165, 1990.

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[20] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[21] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Proc. Interspeech*, 2011, pp. 237–240.

[22] Y. Bao, H. Jiang, L. Dai, and C. Liu, "Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition," in *Proc. ICASSP*, 2013, pp. 6980–6984.

[23] F. Jelinek, "Self-organized language modeling for speech recognition," *Readings in speech recognition*, pp. 450–506, 1990.

[24] G. Fant, *Speech Sounds and Features*. The MIT Press, 1973.

[25] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass, USA: MIT Press, 1998.

[26] J.-L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone latices." in *Proc. Interspeech*, 2004.

[27] P. Matejka, P. Schwarz, J. Cernockỳ, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition." in *Proc. Interspeech*, 2005, pp. 2237–2240.

[28] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc. ICASSP*, 1994, pp. 305–308.

[29] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[30] L. Rabiner, C. Lee, B. Juang, and J. Wilpon, "HMM clustering for connected word recognition," in *Proc. ICASSP*, 1989, pp. 405–408.

[31] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.

[32] A. Martin and C. Greenberg, "The 2009 nist language recognition evaluation," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, 2010, pp. 165–171.

[33] H. Hermansky and S. Sharma, "Temporal patterns (traps) in asr of noisy speech," in *Proc. ICASSP*, 1999, pp. 289–292.

[34] P. Schwarz. (2009) Phoneme recognition based on long temporal context. [Online]. Available: http://www.fit.vutbr.cz/

[35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[36] M. A. Zissman, "Predicting, diagnosing and improving automatic language identification performance." in *Proc. Eurospeech*, 1997, pp. 51–54.

[37] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1, pp. 237–248, 2000.