A Fusion Approach to Spoken Language Identification Based on Combining Multiple Phone Recognizers and Speech Attribute Detectors

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
[2] School of Electrical and Computer Engineering, Georgia Institute of Technology
wyn314@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a fusion approach to spoken language recognition by combining multiple tokenizers with phone and speech attribute models trained on a collection of multilingual corpora with different front-end features. The speech attribute models are trained with bottleneck features extracted from deep neural networks while the phone models are trained with temporal patterns neural network features. By exploiting different combinations of front-end features, fundamental speech units and tokenization models, we demonstrate that speech attribute units are complementary to phone units and produce enhanced performances when they are combined with conventional phone based tokenizers. Tested on the National Institute of Standards and Technology 2009 language recognition evaluation task, leveraged upon diversity in system combination, we find that speech attribute recognition followed by language modeling achieves an additional average relative equal error rate reduction of more than 20% when fused with the state-of-the-art systems with phone recognition followed by language modeling.

Index Terms: spoken language recognition, phonetic features, automatic speech attribute transcription, deep neural network, bottleneck features, phone recognition followed by language modeling, manner and place of articulation

1. Introduction

Spoken language identification (LID) is a process of deciding the language identity of a test utterance. Generally speaking, there are two main LID categories: the acoustic and phonotactics approaches [1, 2, 3, 4]. The acoustic approach extracts discriminative features from the speech signals and employ them to build models, such as Gaussian mixture models (GMMs) [5], to determine the language identity. On the other hand, the phonotactics approach is usually accomplished by phone recognition followed by language modeling (PRLM) [6] to first decode the utterance into a sequence of phones and then an interpolated n-gram language model is used to estimate the probability of the obtained phone streams which is often different for different languages to be identified. It can also be extended to Parallel-PRLM (PPRLM) by incorporating multiple language-dependent phone recognizers and building the corresponding set of language models [3]. Non-phonetic units, such as acoustic segments [7], have also been utilized and produced competitive LID results [8].

In this paper we adopt the phonotactics approach to LID by modeling a set of speech attribute units which is common to all spoken languages [9]. We consider this set of attributes as language universal, and it has been adopted for LID to deliver good LID performances [10] in which the phone recognizer (PR) in PRLM is replaced by a universal attribute recognizer (UAR) focusing on acoustic phonetic features representing manner and place of articulation which were used in a recently proposed automatic speech attribute transcription (ASAT) paradigm [9, 11] for automatic speech recognition (ASR).

We see two main advantages with UAR. First, these units are defined universally across multiple languages [4, 10, 11]. As a result, it alleviates the problem missing phones in the front-end phone recognizer of PRLM systems [3]. It facilitates sharing of speech data from different languages to enhance the modeling capability. Second, the size of the attribute inventory, typical a dozen or so, is smaller than that of various phone inventories, typically a few dozen. In addition, context-dependent (CD) attribute models [12] can also be established to improve the transcription accuracies offered by context-independent (CI) attribute models and therefore enhancing the LID performance, just like in the case of CD-based acoustic modeling in ASR. Popular acoustic phone models, such as hidden Markov models (HMMs, e.g., [13]) and deep neural networks (DNNs, e.g., [14]), can all be adopted to model speech attributes.

To explore the diversity strategy often adopted in ASR [15, 16, 17] in system combination, in this experimental study, we focus our attention on combinations of multiple systems, each based on different units and features. We show that units with different acoustic definitions often exhibit complementary discrimination power such that even different acoustic features and models are adopted for each individual system, the overall performance is often additive when they are combined. Tested on the National Institute of Standards and Technology (NIST) 2009 language recognition evaluation (LRE) task, we find that our proposed fusion framework achieves an additional relative average equal error rate (EER) reductions of more than 20% from that of the state-of-the-art PRLM systems.

2. Universal Speech Attribute Modeling

2.1. Modeling place and manner of articulation

The set of universal speech attributes used in this study is the same as in [18] listed in Table 1, consisting of place and manner of articulation, known as distinctive features [19], commonly adopted to characterize acoustic phonetics [20] of speech sounds for all spoken languages. With these units we can construct universal manner recognizer (UMR) and universal place recognizer (UPR). In addition to these attributes, we also employ the ”silence” token to represent the soundless segments and ”noise” token to represent the noisy background fragments,

978-1-4799-4219-0/14/$31.00 ©2014 IEEE

respectively. Nonetheless the decoded noise tokens are ignored when building language models.

Table 1: *attributes studied in our work.*

Manner	affricate, fricative, nasal, vowel, voiced-stop, unvoiced-stop, glide, liquid, diphthong, sibilant
Place	alveolar, alveo-palatal, dental, glottal, high, bilabial, labio-dental, low, mid, palatal, velar

To build LID acoustic models, shifted delta cepstra (SDC) [5] features together with other popular features, such as perceptual linear prediction (PLP) [21] and temporal pattern (TRAP) [22] features, have been adopted. To train the UAR mentioned above, we utilize bottleneck features (BNF) generated by a structured deep neural network (DNN) [23, 24, 25].

2.2. Context-dependent UAR

We also incorporate context-dependent (CD) models into developing UMR and UPR [18]. For the UMR with 10 units and UPR with 11 units we used in this study shown in Table 1, we can get 121 right-context-dependent (RCD) [26] attribute units of manner and 144 RCD attribute units of place, respectively, on account of the silence token. One thing to note is that the RCD attributes are not all presented in the English and Mandarin data we used in following experiments.

2.3. Complementarity of speech attribute detectors to PRs

In this work, we explore fusion of multiple systems with different speech units. This is inspired by the advantage and complementarity of universal speech attributes to language-dependent phonemes. In the phonotactic approaches to LID, the accuracy of phone recognizer is a critical factor, but not the only one for LID performance. In other words, if a phoneme of another language to be recognized is always recognized as the one in the phone set designed for the phone recognizer, it is fine to model it in the language model based on the assumption of similarity between them. If some phonemes are very different from the phonemes of the language for phone recognizer, they cannot be represented well in language modeling, which is quite common for spoken languages in different language families.

We could alleviate this problem by using attribute units that are potentially language-universal across all spoken languages. Meanwhile, due to the small size of the attribute inventory for manner or place of articulation, a single UAR based LID system may not achieve the comparable performance of a PR based LID system [12]. In this study, we show the complementary nature of speech attribute detectors to phone recognizers by fusing multiple tokenizers with phones and attributes.

3. Description of Fusion Systems for LID

The key idea of phonotactic approaches to LID is to explore the lexical-phonological rules that determine the combinations of different acoustic units [3, 4] in different spoken languages. It usually consists of a tokenizer front-end and a back-end of n-gram models. Using data transcribed with phones and universal speech attributes, we need to build a tokenizer for each set of units to transcribe a given utterance into a sequence of token units first. Then we can build n-gram models based on the chosen units to approximate the probability distribution of the co-occurrences of multiple units. The most common back-end for phonotactic approaches is the n-gram unit language models and vector space models (VSMs) [4]. The n-gram language models [27] describe the distribution as the weighed sum of the probability of different order of n-gram counts. While the VSMs represent the co-occurrences of multiple tokens [8] under the bag-of-unit description [28].

As mentioned above, we build our LID systems based on phonotactic approaches by fusing multiple phone recognizers and speech attribute detectors trained on multilingual databases and different acoustic features. A block diagram of the system is shown in Figure 1. The back-end we used here is the n-gram language model. With n-gram counts generated by the front-end, we can build the language models for speech units. In our study, we have built both PRLM system and universal attribute recognizer followed by language modeling (UARLM) system. The conventional Temporal Patterns Neural Network (TRAPs/NN) [22] based PR is our baseline system which is denoted as TRAP-PR. The PR and UAR based on GMM-HMMs using BNF (BNF-PR and BNF-UAR) are our proposed recognizer. The corresponding LID systems with language model are denoted as TRAP-PRLM, BNF-PRLM, and BNF-UARLM, respectively.

3.1. Multiple speech tokenizers for LID

Multiple tokenizers are designed based on a diversity strategy for system combination. In our BNF-UAR and BNF-PR systems, we need to train a deep neural network as a feature extractor first. The input to this DNN is a PLP based feature vector. The output layer label is the tied-states of tri-phone GMM-HMM models. DNN training is usually split into two steps. First, in the pre-training stage, we create a generative model layer-by-layer with each layer trained as a Restricted Boltzman Machine (RBM) [29]. After all the RBMs are obtained, we stack them to generate the deep belief network (DBN) [29]. Second, we conduct the fine-tuning step in which we employ the back-propagation (BP) algorithm [30] to update the parameters with a minimum cross-entropy criterion [31].

The BNF-UAR is implemented with monophone GMM-HMMs trained with maximum likelihood (ML) criterion [32]. Each attribute is modeled by an HMM with 3 emitting states. Each state has 80 Gaussian mixture components. Here we have built CI-UMR, CI-UPR, RCD-UMR and RCD-UPR. They convert the utterance into transcriptions on different symbols for free attribute decoding with an open attribute loop grammar. Furthermore, we have trained the BNF-PR in the similar way.

3.2. Fusion of the PRLM and UARLM systems

As is shown above, we have developed multiple tokenizers with different acoustic features and models. With them we can accomplish the decode procedure to get parallel token sequences. After that, we can build independent sets of language models and employ them to estimate the probability of n-gram co-occurrences. Finally, we fuse the outputs of all systems to obtain the final score which is used to make the LID decision.

As for the fusion, we combine the output scores of each UARLM and PRLM system through the Gaussian back-end [33, 34]. Assuming that we have N tokenizers and the target language number is M, then we can get a vector of $M \times N$ dimensions as the output score of a test utterance, which is the input to train the Gaussian back-end. During the training phrase with the development set, we train M Gaussian models with diagonal matrices under the MLE criterion for the corresponding

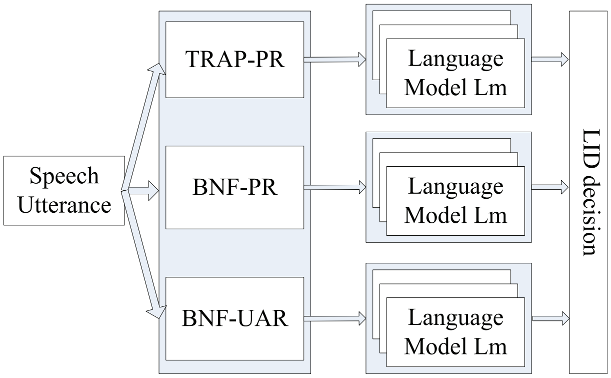

Figure 1: *Block diagram of our fusion system*

target languages [33]. In the evaluating stage, for a test utterance, the Gaussian models can yield the likelihood vectors with each element corresponding to each target language. The one with the highest likelihood is chosen as the target language.

4. Experimental Results and Analysis

4.1. Experimental setup

To verify the performance of our bottleneck feature-based UAR and PR, we conducted our LID experiments on LRE09 data set which included 23 target languages [35]. The utterances in the training set were recorded with two channels, namely Conversational Telephone Speech (CTS) and narrowband Voice of America (VOA) [35]. As the training data was not balanced, we only made use of a subset of 15 hours speech data for each target language. And we split them 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 lengths for the test utterances, i.e., namely 3s, 10s and 30s.

For training of the DNN as the BNF extractor, we employed the 309-hour Switchboard-I training set [36] and the in-house corpus of Mandarin telephone conversational speech of 1000 hours, respectively. The feature to feed into the DNN was the 43-dimensional feature vector consisting of 13-dimensional PLP feature 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. The frame length was 25 msec with a frame shift of 10 msec. Here we trained the DNN with 5 hidden layers. There were 55 units for the bottleneck layer in the middle of hidden layers and 2048 units for other hidden layers. As for labels of DNN output, which were produced by the forced-alignments with the triphone GMM-HMMs, we used 9004 tied states and 6004 tied states for models trained with the English Switchboard corpus and Mandarin corpus, respectively.

4.2. PRLM and UARLM Results

With the listed attributes inventory in Table 1, we have developed both RCD-UMR and RCD-UPR together with CI-UMR and CI-UPR. Figure 2 plots the DET curves [37] for the performance of both UMR and UPR under CI and RCD circumstances on the 30s-length LRE09 task in which we can observe that the RCD-UAR performed superior to CI-UAR. Moreover, it can be seen that the EER increased from 9.84% to 12.75% when switching from the CI-UMR to CI-UPR system while the EERs were nearly the same under the RCD circumstances. This reduction is because of the closer similarity for place of articulaiton we discussed in Table 1. For example, the pairs of alveolar and alveo-palatal, dental and labio-dental are more similar to each other than for other attributes. This result indicates that an accurate decoding is a key to improving attribute based LID systems. Other techniques proven to improve ASR accuracies can also be adopted.

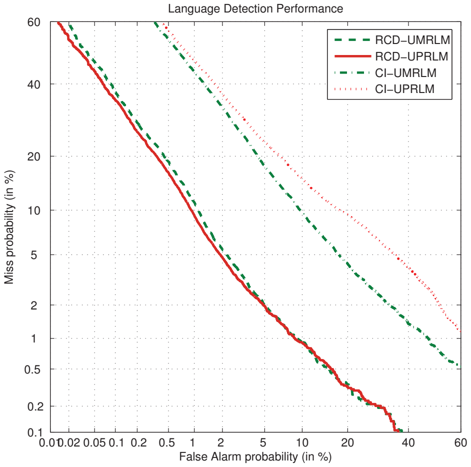

Figure 2: *DET curves for different UARs on 30s task of LRE09.*

Table 2: Performance (EER and C_{avg} in %) comparison between PRLM and UARLM systems on the LRE09 tasks

	30s		10s		3s	
	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
BNF-PRLM with EN: P1	2.88	2.87	6.62	6.6	19.6	19.43
RCD-UARLM Fusion: UC1 (UMR+UPR)	2.59	2.58	6.11	6.03	17.51	17.43
BNF-PRLM with MA: P2	3.08	3.03	7.79	7.78	21.93	21.65
RCD-UARLM Fusion: UC2 (UMR+UPR)	2.98	2.92	7.52	7.48	21.66	21.53
TRAP-PRLM with RU: P3	2.42	2.4	6.42	6.38	18.92	18.70
TRAP-PRLM with HU: P4	2.62	2.62	6.65	6.62	18.88	18.82

Table 2 lists performance comparisons of the two styles of TRAP based PRLM systems and our BNF based PRLM and UARLM systems. First we utilized the two top-performance phone recognizers, P3 and P4, developed with the temporal patterns neural network (TRAPs/NN) scheme for Russian (RU) and Hungarian (HU) [38], respectively, by Brno University of Technology (BUT). Both P3 and P4 have been widely adopted as benchmark tools for LID system combinations and performance comparisons because they have been known to deliver excellent LID results. We will do the same next.

In Table 2, BNF-based phone recognizers include an English PR (P1) trained with the Switchboard corpus and a Mandarin PR (P2) trained with the aforementioned corpus. From these results, for example for 30sec test utterances, we found, not surprisingly, that using P1 (EER of 2.88%) and P2 (EER of 3.08%) alone was not as competitive to P3 (EER of 2.42%) or P4 (EER of 2.62%). However the bottleneck features were shown to be effective under the GMM-HMM framework for phonotactic approaches to LID. UC1 and UC2, in Table 2, were based on combining English and Mandarin RCD-UMRs and RCD-UPRs, respectively. Moreover, our UARLM systems UC1 (EER of 2.59%) and UC2 (EER of 2.98%) alone delivered lower EERs than those for P1 and P2 when they were based on the same features. They also gave competitive performances to the TRAP/NN based PRLM systems, indicating that our attribute recognizers based on manner and place of articulation worked as well as phone recognizers. Other results for utterance lengths of 10s and 3s showed similar trends.

Table 3: Fusion results on LRE09 in EER(%) and $C_{avg}(\%)$

	30s		10s		3s	
	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
Fusion: F1 UC1+P1	2.28	2.26	5.08	5.05	15.77	15.70
Fusion: F2 UC2+P2	2.23	2.19	5.13	5.10	16.86	16.81
Fusion: F3 P3+P4	1.78	1.78	4.70	4.65	15.24	15.15
Fusion: UF3 UC1+UC2	2.16	2.12	4.39	4.35	14.78	14.67
Fusion: UC1+F3	1.59	1.58	3.49	3.47	12.11	12.06
Fusion: UC2+F3	1.61	1.60	3.80	3.75	13.38	13.29
Fusion: FALL UC1+UC2+F3	1.55	1.55	3.26	3.26	11.29	11.27

4.3. Fusion Results

Table 3 presents all the fusion results on the evaluate set of the LRE09 task. F1 is the fusion of UC1 and the English PRLM system P1 while F2 is the fusion of UC2 and Mandarin PRLM system P2. F1 achieved an average relative EER reduction of 20% when compared with P1 even though P1 and UC1 were derived from bottleneck features and trained with the same speech corpus. Similarly, F2 improved the performance of P2 by 27%, 34% and 23% relatively for 30s, 10s and 3s tasks, respectively. These clearly verifies our conjecture that the tokenizers based on the universal speech attributes is complementary to the conventional phone recognizer for spoken language recognition.

Next in Table 3 we use F3 to denote fusion of the Russian and Hungarian PRLM systems, and UF3 to denote fusion of the English UARLM system UC1 and the Mandarin UARLM system UC2. Although UF3 (at EER of 2.16%) performed inferior to F3 (at EER of 1.78%) for the 30s task, UF3 was slightly better than F3 for the 10s (at EER of 4.39% vs. 4.70%) and 3s (at EER of 14.78% vs. 15.24%) tasks. This indicates that BNF is more potential than TRAP features for more difficult short time task. When fusing F3 with UC1 (at EER of 1.59%) or UC2 (at EER of 1.61%) as listed in Table 3 we obtained some additional performance gains. Finally, we fused UC1, UC2 and F3 together and obtained the best LID EER results of 1.55%, 3.26% and 11.29% for the 30s ,10s and 3s test utterances, respectively. These improvements, especially for short duration utterances, may benefit from combining different acoustic model architectures and various features we used here in addition to the complementary nature between fundamental speech attributes and phones as indicated earlier.

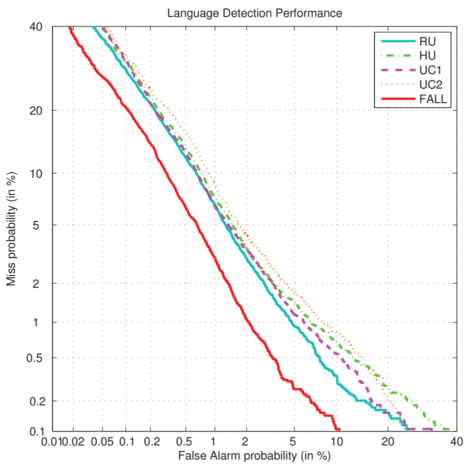

Figure 3: *DET curves for UARs and PRs on 30s task of LRE09.*

Finally, in Figure 3 we plot the DET curves of the fusion results for the aforementioned UARLM systems and the PRLM systems on the 30s tasks. It is clear that fusion of diverse systems always produces improved results. The wider gap in the lower right part than the upper left part of the figures seemed to indicate the false alarm rates could be reduced more than the miss detection rate when speech attributes are integrated into the conventional PRLM systems. In essence, by fusing multiple phone recognizers and our proposed UARs (e.g., FALL in Table 3 and Figure 3) we have obtained a significant performance improvement on the LRE09 task. Nevertheless, the UARs trained on only a limited set of languages could not yet fully reflect the language-universal characteristics of speech attributes as intended in this study. More research is needed.

5. Conclusion

We propose a UARLM framework within the phonotactic approaches to LID. We train multiple right-context-dependent UARs from multilingual databases to construct a diverse family of LID systems based on various sets of models, features and tokenization units. Tested on the NIST LRE09 tasks we found the proposed UARLM system outperforms the conventional PRLM systems using the same front-end features and back-end language classifiers, but only with different tokenization units. The proposed UARLM systems are found to be competitive with the the state-of-the-art PRLM systems. By fusing the proposed UARLM systems trained on the bottleneck features with the best TRAP/NN based PRLM systems, leveraging on the diversity properties required in system combination, we also obtain further performance gains over the PRLM systems in all three test utterance lengths of the NIST LRE09 tasks. The EER reduction is most significant at 26% for the cases with the shortest 3-second test utterances.

6. 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).

7. References

[1] 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.

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

[3] 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.

[4] 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.

[5] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features." in *Proc. Interspeech*, September 2002.

[6] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc ICASSP-94*, vol. 1, April 1994, pp. 305–308.

[7] C.-H. Lee, F. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *Proc. ICASSP*, April 1988, pp. 501–541.

[8] 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.

[9] 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.

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

[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, 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.

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

[14] 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.

[15] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. ASRU*, 1997, pp. 347–354.

[16] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*.

[17] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted rover and cnc." in *Interspeech*, 2006, pp. 537–540.

[18] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

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

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

[21] 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.

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

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

[24] 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.

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

[26] 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.

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

[28] G. Salton, *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

[29] 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.

[30] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[31] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

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

[33] M. A. Zissman, "Predicting, diagnosing and improving automatic language identification performance." in *Proc. Eurospeech*, vol. 1, September 1997, pp. 51–54.

[34] 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.

[35] A. Martin and C. Greenberg, "The 2009 nist language recognition evaluation," in *Proceedings of Odyssey*, 2010, pp. 165–171.

[36] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1, Mar 1992, pp. 517–520.

[37] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eurospeech*, vol. 4, 1997, pp. 1895–1898.

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