AUTOMATIC LANGUAGE IDENTIFICATION WITH PERCEPTUALLY GUIDED TRAINING AND RECURRENT NEURAL NETWORKS

Jerome Braun and Haim Levkowitz

Computer Science Department, University of Massachusetts Lowell
One University Avenue, Lowell, MA 01854, USA

ABSTRACT

We present a novel approach to Automatic Language Identification (LID). We propose *Perceptually Guided Training (PGT)*, a novel LID training method, involving identification of utterance parts which are particularly significant perceptually for the language identification process, and exploitation of these *Perceptually Significant Regions* (PSRs) to guide the LID training process. Our approach involves a Recurrent Neural Network (RNN) as the main mechanism. We propose that, because of the long-range intra-utterance acoustical context significance in LID, RNNs are particularly suitable for the LID task. Our approach does not require phonetic labeling or transcription of the training corpus. LIREN/PGT, the LID system we developed, incorporates our approach. Our LID experiments were on English, German, and Mandarin Chinese, using the OGI-TS corpus.

1. INTRODUCTION

Automatic Language Identification (LID) is the task of identifying the natural language used in a monolingual speech excerpt, in a speaker-independent manner. Approaches to LID include (e.g., [9]) those based on HMMs, VQ and histograms, phonotactics, prosody, as well as techniques utilizing Large Vocabulary Continuous Speech Recognition (LVCSR), and feedforward (MLP) neural networks. We are concerned with the *essential* LID task, the term we use for the LID capability that does not rely on speech recognition capabilities at the word-level and above.

We present a novel approach to the essential LID. *Perceptually Guided Training (PGT)*, a novel paradigm we propose, involves identification of utterance parts which are particularly significant perceptually for the language identification process, and exploitation of these *Perceptually Significant Regions* (PSRs) to guide the training process. Thus, exploitation of the non-uniform distribution of information specific to the language identification process, locating the utterance portions where the levels of such information are elevated (PSRs), and utilizing them to improve the LID training process, are the underlying concepts of the PGT. Our approach involves the Recurrent Neural Network (RNN) architecture as the fundamental LID mechanism [1,3]. We believe that the long-term intra-utterance context plays a major role in the essential LID. The inclusion of this context is our principal motivation for the RNN-based approach. Our approach does not require phonetic labeling or transcription of the speech corpus. The developmental and experimental aspects of our research include a system implementing our approach, LIREN/PGT (**L**anguage **I**dentification with **RE**current **N**eural networks and **P**erceptually **G**uided **T**raining). The LID training experiments with LIREN/PGT show the efficacy of our approach. Our research includes investigations of a number of issues in RNN LID training, and proposes a number of algorithmic solutions for the RNN training for LID. Our LID experiments were on English, German, and Mandarin Chinese, using the OGI-TS [5] speech corpus.

2. RNN-BASED APPROACH

We suggest that the acoustical context range is particularly extensive and important in LID (compared to, e.g., a phonetic recognition task). We propose that the implicit inclusion of the past, inherent in RNNs due to their feedback connections, can account for that context, making RNNs particularly suitable for LID. Regarding the type of input, although LIREN/PGT has provisions for possible future experiments with speech units (in particular the automatically derived fenonic units [3]) its baseline form relies on the acoustic domain feature vectors. The principal reason for this is the desire to deliver to the RNN main engine the maximum of the information available in the signal. A conversion of the acoustic domain features into speech units simplifies the network subsequent operation (a simpler classification problem), but it filters the information, narrowing its scope to the chosen speech unit domain. In the absence of the fundamental understanding of the essential LID process, we chose to attempt the RNN recognition using *all* available information, i.e., the acoustic feature vectors, in spite of the difficulty issue (feature space dimensionality). Our results show that, with appropriate algorithmic provisions, such LID training can be successful.

The global architecture of the LIREN/PGT system is shown in Figure 1. The preparatory stages (B), aided by auxiliary utilities (F) generate the speech data repository contents in the "LIREN/PGT native" form. The VideVox facility [2] (C) is responsible for supplying the PGT information (discussed later) that, together with the corresponding speech waveform data constitutes the LIREN/PGT main speech data repository (D). The feature generation stages (E) produce the feature vector sets (G), the input to the RNN engine. The LIREN/PGT system was developed in C and MS Visual C++ ver. 5.0, on a WindowsNT, 200 MHz Pentium Pro system with 64MB RAM.

Recurrent Architectures in LIREN/PGT. Two related RNN architectures have been used in LIREN/PGT. The first, a *fully* interconnected two-layer recurrent architecture, is shown in Figure 2a. This diagrammatic representation does *not* show the multitude of nodes and connections. The output of *each* node in layer L1 connects to *each* node of layer L2 ((A) is a fully connected bipartite graph). State input nodes receive their inputs through the delay stage from the state output nodes in layer L2. The external output nodes deliver the language probability (one node per language). The number of state nodes impacts the degree of context inclusion by the network. In this work the number of state nodes was often about 160. With 21 input nodes, a bias node and two external output nodes the network contains a total of 344 nodes, and 118336 weight connections. Neural networks of such size present serious challenges in development and training (further exacerbated by the delay $/z^{-1}/$ loops which, during training, turn the network into a large multilayer structure, replicating in time the above spatial view). The second neural network architecture in LIREN/PGT, shown in Figure 2b, is RTRL [8]. The L1 nodes consist of bias and input nodes, and are fully interconnected to layer L2 via the connection structure (A). Thus (A) contains $(b+i)(r-t)$ links, where b, i, r, and t, are the number of bias, input, RTRL and target nodes, respectively. All nodes in layer L2 are also connected through the delay stage (z^{-1}) to all nodes of L2. Thus the feedback loop involves a fully interconnected set of links (B, C). As in the previous case, our diagrammatic representation hides the huge size of the network (Figure 2b does not show explicitly the multitude of nodes and connections).

Target functions. In LID, the RNN target function definition issue is not trivial. While the input behavior (speech) is quite dynamic, the output (language) remains static upon the change of speakers, their speaking style or gender, and changes *only* at those utterance boundaries where the two utterances, adjacent in the training data stream, belong to *different languages*. In [1,3], we have described three target function models of LIREN/PGT: the ***piece-wise constant***, the ***exponential rise***, and the ***linear fuzzy half*** models. The *linear fuzzy half* model we proposed, smoothens the training process by introducing overlaps in language probabilities during the first half of each utterance: while the true-language target vector component increases, the components (probabilities) for other languages decrease correspondingly. This fuzzy overlap stops at mid-point of the utterance, and the target components remain constant across utterances, until a language change occurs, at which point the process repeats. The fuzzy overlap effect starts only on those utterance boundaries on which the language *actually* changes; other boundaries have no effect (regardless of, e.g., speaker or speech mode changes between utterances). The "fuzzy half" target function resulted in the best performance.

Acoustic Signal Processing. LIREN/PGT includes three types of acoustical processing front-end subsystems, for a comparison of different feature vector types' efficacy in RNN LID. In the *spectral domain based feature vector subsystem*, similar to [6], short term Fourier analysis (FFT) is performed, and the power spectrum is divided into twenty mel-scale bins. The resulting twenty spectral energy coefficients and the signal energy form an $R^{(21)}$ feature vector space. In the *cepstral domain-based feature vector subsystem*, the 16th order LPC based cepstral coefficients and, as in [6], the fundamental frequency estimate and the voicing level, form an $R^{(19)}$ feature vector space. In the *RASTA-PLP based feature vector subsystem*, the 8th order RASTA-PLP [4] coefficients and the signal power form an $R^{(9)}$ feature vector space. Our experiments, comparing LID performance of these three feature vector types, did not show major performance differences between the lower-dimensional (9D) RASTA-PLP feature vectors and substantially higher-dimensional (e.g., 21D) spectral or cepstral feature vectors. These results indicate that RASTA-PLP based features are particularly suitable for LID.

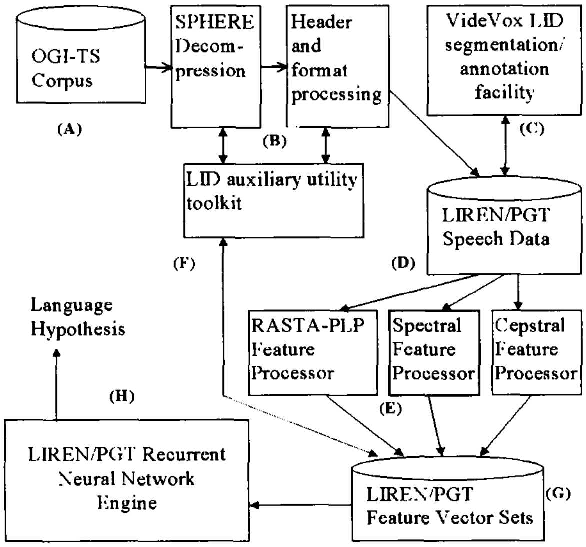
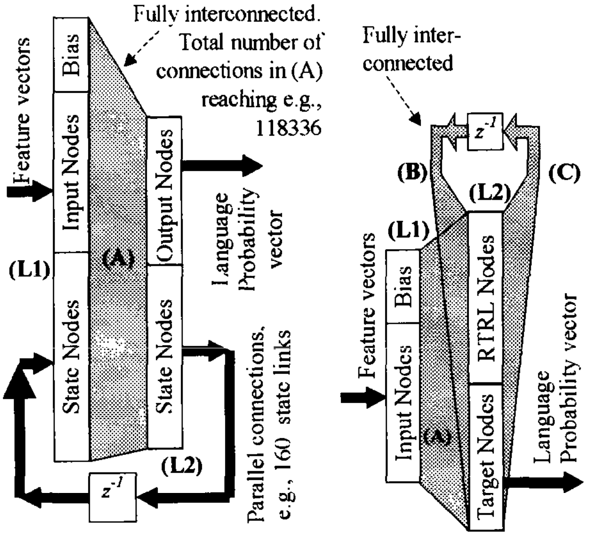

OGI-TS Corpus
(A)
SPHERE Decom-pression
Header and format processing
(B)
VideVox LID segmentation/ annotation facility
(C)
LID auxiliary utility toolkit
LIREN/PGT Speech Data
(F)
(D)
Language Hypothesis
RASTA-PLP Feature Processor
Spectral Feature Processor
Cepstral Feature Processor
(H)
(E)
LIREN/PGT Recurrent Neural Network Engine
LIREN/PGT Feature Vector Sets
(G)

Figure 1: LIREN/PGT global architecture



(a) 2-layer Recurrent Neural Network

(b) RTRL Recurrent Neural Network

Figure 2: LIREN/PGT recurrent neural network architectures

3. PERCEPTUALLY GUIDED TRAINING

Perceptually Guided Training (PGT), a novel method for LID training, is a key component of our overall approach. Perceptual experiments indicate [3] that humans notice certain specific utterance parts when listening to an unknown language. We propose that language-identity-specific information is distributed in a non-uniform fashion along the utterance, and that human listeners are able to spot the locations within the utterance where the levels of such information are elevated. We do not know what aspects of the total information present in the speech signal actually constitute the language-identity-specific information. However, we propose that the location of utterance portions that are "characteristic" for human listeners, can be determined experimentally, and that these regions exhibit elevated levels of the language-identity-specific information. In PGT, we perform the detection of such regions and, using the detection results, we subsequently guide the training process to emphasize these parts of the training data (utterances) that are particularly language-identity-significant perceptually.

3.1 PSR Identification

We use VideVox, a dedicated facility we developed [2], to identify regions within the utterance that appear to listeners as particularly characteristic of the language used in the given utterance. The data obtained in this process represent what we termed as the *Perceptually Significant Regions (PSR)*, defined as the regions where the density of language-identity-specific information is particularly high [3]. The PSR boundaries are not required to be precise and are described probabilistically. We do not constrain PSRs in any way, e.g., we do not impose any assumptions or constraints on their duration or content.

PSR identification is accomplished through perceptual experiments, in which human subjects interact with VideVox. During the experiment session, VideVox is used [2] to present the utterances in different languages to the subject. VideVox facilities [2], designed specifically for the PSR identification task, allow subjects to identify and designate PSRs. We performed perceptual experiments to elicit the PSR data for English, German, and Mandarin Chinese (subjects interacted with VideVox, using its facilities to identify PSRs). Our experiments indicate that the PSR detection is possible. The subjects had no difficulty in identifying "characteristic" regions for the languages they did not know. The overlaps between the areas identified by the subjects indicate the existence of PSRs common to different listeners. In addition, we observed the following trends during the PSR identification. For English, the PSRs often included phrases containing /th/, /dh/, /r/, "the", /jh/, and vocal pauses filled with /ae/, /a/. For German, PSRs often included phrases containing *ich*, *ich-bin*, *auf*, *che* (as in *kuche*), *ro* (as in *buero*), *ein*, *die*, *gen* (as in *gegangen*) *den* or *ten* (as in *arbeiten*), and "*stra*" (as in *strasse*). For Mandarin Chinese, PSRs often included rapid successions of syllables generally starting with the Mandarin Chinese phones similar to /y/, /ch/, and /sh/. However, while the above trends appeared to represent expectable phonetically oriented patterns (e.g., /th/ in English), others were relatively unexpected (e.g., "*ro*" for German). It should be emphasized that, while the above phonetic sequences were observed in PSRs, the actual PSRs that contained them were typically much longer in duration, and thus should not be considered as chiefly due to, or reducible to, these phonetic sequences. We believe that PSRs arise from effects that are *not* restricted only to the phonetic domain and that they include long-range (beyond phonetic level) phenomena.

Based on our experiments, it appeared that in 10-sec. utterances (OGI-TS) the subjects could identify up to four PSRs. After that, the efficiency dropped off visibly and the subjects were likely to "fix" on a specific short interval within the utterance. If confirmed, this phenomenon may also be related to our procedure, e.g., the repetitive listening to the entire utterance (allowed by VideVox), to which subjects often resorted after concluding the identification of a PSR. Regarding the utterance duration, the PSR identification appeared to be easier in longer utterances (story-bt), than in short (10 sec.) ones.

3.2 PSR re-exposure algorithm

We developed two techniques by which the PGT makes use of the PSR data during the neural network training. The first technique, the *target function profiling* algorithm, involves a dynamic modification of the target function at the PSR locations, reinforcing the correct language target function vector component at those locations. The second technique, the *PSR re-exposure* algorithm, proved to perform better of the two. The essence of the PSR re-exposure algorithm is a controlled and automatic increase of the neural network's exposure during training to the language-identity-specific information contained in the PSRs. Each PSR is presented (re-exposed) to the network r times in each training epoch (r is an adjustable re-exposure factor). When LIREN/PGT operates in the PSR re-exposure training mode, the feature vector stream processing includes the detection of the PSRs' presence. The boundaries of each PSR are determined and the PSR is in effect replicated from the training viewpoint. Thus the influence of the PSR areas on the training process is reinforced r times. The PGT intensity [3] is proportional to r. Our experiments showed that $r=8$ was a good operating point for the algorithm. This operating point appeared to be essentially independent of the PSR duration and specifics.

We performed a large number of experiments to determine the efficacy of the PGT method. The English/German experiments were particularly interesting, given the linguistic proximity of the two languages. Our experiments showed a consistent superiority of the PGT training performance over a non-PGT training, with an improvement of about 9%.

4. RNN TRAINING IN LID

4.1 Backpropagation Through Time

Backpropagation Through Time (BTT) [7] was one of two training modes used in this work. We used the adaptive weight update algorithm [6]. Our initial LID experiments with it showed convergence difficulties, attributed to the difficulty of the LID task directly based on acoustic features. Among the modifications we introduced in BTT to achieve training convergence, the step size range restriction, and the avoidance of near-zero gradient operations, were found most effective [1]. Limiting the step sizes (in each weight space dimension) to a band of four orders of magnitude of the initial step size helped prevent step size drifting. When the weight update decision (gradient sign test [6]) originated from near-zero values, i.e., the local gradient $|\partial J(t)/\partial w_{ij}|$ was below a low-value threshold, the corresponding weight update was withheld. This had a stabilizing effect on the training process.

Number of state nodes. The ability of the RNN to account for the context (past) is influenced by the number of state nodes. We studied the effect of the number of state nodes on the LID training. RNNs with 75 nodes in the state layers performed acceptably well. Our best results involved 120 to 160 state nodes; we used 160 state nodes in many of our experiments.

BTT sequence extent. The BTT sequence (expansion) length is the extent of the *explicit* context inclusion in BTT. We experimented with two types of context interval: a selectable fixed length, and a variable length equal to the length of utterance. The fixed BTT sequence lengths below 0.3 utterance length performed relatively poorly, reflecting an insufficient context inclusion level. A more complex, variable sequence length approach performed better, while the best performance was attained with fixed sequence lengths in the range of 0.3 to 0.9 utterance lengths.

4.2 Aperiodic Update Recurrent Training

The second training algorithm in LIREN/PGT is the Aperiodic Update Recurrent Training (AURT) algorithm we developed. AURT is in essence a modified Williams-Zipser RTRL algorithm [8]. The AURT method is based on the specific character of the LID task: the input feature vectors are naturally grouped by utterances. Considering this, and the possibilities of changes (speaker, speech characteristics, or language) from one utterance to another, we tie the reset of the RTRL impact coefficients (representing the influence of *any* weight on the output of *any* node) to the utterance *endpoints*. Thus, unlike in RTRL, the impact matrix reset mechanism in our method is *aperiodic*, since it is dependent on the (variable) utterance lengths. In AURT, both the aperiodic reset of the impact matrix, and the weights update process take place at those points. We performed a multitude of LID training experiments, on English, German and Mandarin Chinese, with and without PGT, to determine the efficacy of AURT. The non-PGT AURT training typically exhibited a convergent but slow learning behavior. The PGT non-AURT experiments often exhibited rapid learning, especially in the initial phases. However, not infrequently, we observed a major instability developing later on, followed by divergence. On the other hand, the PGT AURT, under the same conditions, offered instability-free convergence. Based on these experiments, we conclude that the main advantage of the AURT algorithm is its stability, while preserving an acceptable learning progress rate.

5. SUMMARY

We have described a novel approach to Automatic Language Identification (LID), involving *Perceptually Guided Training (PGT)* and Recurrent Neural Networks. PGT is based on locating and utilizing *Perceptually Significant Regions* (PSRs), the utterance regions that are particularly significant perceptually for the language identification process. We demonstrated that PGT improves the LID training performance, with consistent improvements of around 9% versus the non-PGT mode. We suggested that intra-utterance acoustical context is particularly critical (and extensive) in LID. We proposed Recurrent Neural Networks as the central identification architecture, motivated by a capability of an implicit inclusion of that context via the feedback mechanism of the RNN. Our approach does not require phonetic labeling or transcription of the training data. We proposed several algorithmic directions for RNN training in LID. These include the Aperiodic Update Recurrent Training (AURT), and LID-related modifications to the Backpropagation Through Time. We have shown experimentally the convergence and the feasibility of RNN training for LID, when the algorithmic solutions we have proposed are employed. The LIREN/PGT system we have developed implements our approach to LID. Experimental aspects included LID experiments with English, German, and Mandarin Chinese, using the OGI-TS speech corpus. The results of our experiments demonstrate the promise of Perceptually Guided Training, and Recurrent Neural Networks, for Automatic Language Identification.

6. REFERENCES

1. Braun, J., and Levkowitz, H., "Automatic Language Identification with Recurrent Neural Networks," Proc. IJCNN, 1998
2. Braun, J., and Levkowitz, H., "Internet Oriented Visualization with Audio Presentation of Speech Signals," *Proc. SPIE Conf. Visual Data Exploration and Analysis*, Jan. 1998.
3. Braun, J., "Automatic Language Identification with Recurrent Neural Networks," *Sc.D. Dissertation*, University of Massachusetts, Lowell, 1997.
4. Hermansky, H., and Morgan N., "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, 1738-1752, October 1994.
5. Muthusamy, Y.K., Cole, R.A., Oshika, B.T., "The OGI Multi-language Telephone Speech Corpus," *Proc. ICSLP 92, Banff, Alberta, Canada*, October 1992.
6. Robinson, T., "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Trans. Neural Networks*, vol. 5, no. 2, March, 1994.
7. Werbos, P.J., "Backpropagation through Time: What it does and How to do it," *Proc. IEEE*, vol. 78, 1550-1560, October 1990.
8. Williams, R.J., and Zipser, D., "A learning Algorithm for Continually Running Fully Recurrent Neural Networks," *Neural Computation*, vol. 1, 270-280, 1989.
9. Zissman, M., "Overview of Current Techniques for Automatic Language Identification of Speech," Proc. IEEE Automatic Speech Recognition Workshop, 1995.